UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

IMPROVING WOFS-PHI WATCH TO WARNING SEVERE WEATHER GUIDANCE THROUGH NEW PREDICTOR AND TARGET DATASETS

A THESIS
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
MASTER OF SCIENCE

By

RYAN MARTZ
Norman, Oklahoma
2025

IMPROVING WOFS-PHI WATCH TO WARNING SEVERE WEATHER GUIDANCE THROUGH NEW PREDICTOR AND TARGET DATASETS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Adam Clark (Chair)

Dr. Eric Loken

Dr. Amy McGovern

Dr. Aaron Hill

Acknowledgments

This thesis would not have been possible without the guidance and support from many over the years. First, I would like to thank my advisor, Dr. Eric Loken for being a spectacular mentor. He was always available to meet whenever I stopped by his office, and he would respond to any question or concern promptly even when it was outside of working hours. Despite his busy schedule leading a large grant and planning multiple hazardous weather testbed experiments each year, he always made time to discuss my work and provide thoughtful feedback and never got frustrated with my mistakes that set me back. His encouraging guidance inspired me and gave me the confidence to decide to pursue my PhD after finishing my master's degree.

Second, I thank Drs. Adam Clark, Amy McGovern, and Aaron Hill for their guidance and feedback as a part of my committee. I am appreciative of the time that they took out of their schedules to meet with me about my work and to provide valuable feedback on my thesis.

Third, I thank the other scientists within CIWRO/NSSL that have provided guidance along the way. Specifically, I thank Drs. John Cintineo and Montgomery Flora for their technical guidance and feedback on my methods along the way.

Fourth, I would like to thank my past mentors who shaped me and my interests prior starting my master's program. Specifically, I would like to thank Dr. Adam Houston for his three years of reserach mentorship during my undergraduate years and two further years of guiding me through the review process for my first journal article publication. I would also like to thank Mr. Mark Klein, Mr. Greg Carbin, and Ms. Allison Santorelli for their guidance during my Lapenta Internship program that helped me to decide to pursue a graduate degree. I would not be where I am today without the guidance of these four mentors.

Lastly I would like to thank my parents, Andrea and Erik Martz, and my brother, R.J. Martz for their support and unconditional love. My parents helped nurture my

interest in meteorology and provided extensive help to ensuring I could pursue this passion in my undergraduate career and beyond. R.J. has been the best older brother that I could ask for as a lifelong best friend. Just as I would not be where I am without the guidance of my mentors, I would not be who I am without the support and love from my family.

This work was supported by NOAA/WPO Weather Testbeds grant #NA22OAR4590530 and by NOAA-University of Oklahoma Cooperative Agreement #NA21OAR4320204.

Table of Contents

List Of Tables

List Of Figures

Abstract

Warn-on-Forecast System-Probabilistic Hazard Information (WoFS-PHI) is a real-time machine learning algorithm that predicts individual severe weather hazards (hail, wind, and tornadoes) for up to 4 hours of lead time. The original version of WoFS-PHI used predictors from WoFS and ProbSevere Version 2 (PS2) to predict local storm reports (LSRs). While previous work found WoFS-PHI to be more skillful than machine learning baseline forecasts from WoFS or PS2 individually, an open question is whether using different predictor and/or target datasets would improve the skill of WoFS-PHI.

In this thesis, multiple experiments are conducted to examine the (combined and individual) influence of: ProbSevere Version 3 (PS3) and Tornado Probability Algorithm (TORP) predictors, 1-h vs. 2-h time windows, and hazard-specific warning machine learning targets. New WoFS-PHI models were trained with varying sets of predictors and target datasets. Models were verified using Brier Skill Score, performance diagrams, and attributes diagrams. Predictor importance was analyzed through relative tree interpreter and accumulated local effect curves.

Results show that including PS3 and TORP in the predictors modestly improves forecast skill for all three hazards, but especially for tornadoes. Predictor importance results showed that PS3 predictors were slightly more important to WoFS-PHI than PS2, and PS3 were less important for 2- than 1-hour forecasts regardless of lead time, but WoFS predictors were more important for 2- than 1-hour forecasts. TORP had relatively minimal impacts compared to WoFS and ProbSevere due to a lack of TORP objects in the training dataset. However, when TORP objects were present, they were quite influential to forecasts of all hazards. Results showed that BSSs for WoFS-PHI models trained on warnings and LSRs were much higher than BSSs for WoFS-PHI models trained on LSRs only. This suggests that WoFS-PHI is more skillful at predicting warnings relative to warning climatology than predicting LSRs

relative to LSR climatology. In other words, warnings-and-LSRs can be a more predictable target for WoFS-PHI than LSRs alone. While additionally training on warnings offers both advantages and disadvantages, overall findings here suggest that warnings should at least be considered to supplement LSRs as a target for severe weather forecasting.

Chapter 1

Introduction

1.1 Introduction

During the late 2000s and early 2010s, the National Oceanic and Atmospheric Association (NOAA) began developing a new forecasting paradigm known as Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018). The goal of FACETs is for the National Weather Service (NWS) to be able to issue probabilistic forecast products for high-impact weather events with continually updating spatial and temporal specificity (Rothfusz et al. 2018). While the FACETs paradigm applies to all types of hazardous weather, one particular challenge has been noted regarding information voids that arise within the forecasting period between watch and warning issuance (often 0-4 hours of lead time; Rothfusz et al. 2018). This is known as the watch-to-warning period.

The watch-to-warning period requires special consideration due to the large variation between the spatiotemporal scales of watch and warning forecasting. Watches often having a geographic scale of $20000 - 40000$ mi^2 and time scale of $6 - 8$ hours (Storm Prediction Center 2017), while individual storm warnings cover select counties and are issued on the timescale of minutes instead of hours (Rothfusz et al. 2018). As the spatiotemporal scales of predictability of severe hazards increases during the watch-to-warning period, continuously updating probabilistic forecast products allow for a variety of end user needs to be met across lead times and

decision thresholds (Morss et al. 2010; Hoekstra et al. 2011; Senkbeil et al. 2013). However, it has been shown that forecasters struggle to develop such probabilistic forecast products without "first guess" guidance fields (Stumpf et al. 2008). Some probabilistic guidance tools have been created for the 0-4 hour watch-to-warning period, such products generate probabilities from applying machine learning techniques to NWP output (Flora et al. 2021; Harrison et al. 2022; Hempel 2022). However, even with rapid data assimilation, NWP models struggle to initialize storms with the correct location and intensity which can be problematic for severe hazard forecasting in the early portion of the watch-to-warning time scale (Guerra et al. 2022). While some machine learning products that combined observation-based data with NWP output have been developed to address these issues, they either did not focus on severe hazard forecasting (Williams et al. 2008) or did not cover the timescale of watch-to-warning forecasting (Schmidt et al. 2024).

To this end, a new watch-to-warning forecast product was developed in 2023 to explicitly provide probabilistic guidance in the 0-4 hour time span for severe hail (≥ 1 inch in diameter), wind (≥ 58 mph gusts), and tornadoes called Warn-on-Forecast Probabilistic Hazard Information (WoFS-PHI; Loken et al. 2022b, in review). This machine learning product attempts to address the limitations of NWP by merging NWP data with now-casting guidance to capture both near term storm evolution and multi-hour NWP forecasts in one probabilistic product (Loken et al. 2022b, in review).

The purpose of this thesis is to first understand if changes to the predictor and target datasets of WoFS-PHI can improve its probabilistic severe hazard guidance. Second, this thesis will aim to explain how the new predictor and target datasets are driving changes in WoFS-PHI forecasts.

1.2 Research Background

1.2.1 Convection Allowing Models/Ensembles (CAMs/CAEs)

The first NWP models to allow for convective forecasts used convective parameterizations rather than true simulations of convection due to computational limitations (Manabe et al. 1965). While the first successful attempt at three dimensional modelsing of storm scale processes was achieved shortly after in the 1970s (Brooks et al. 2019), it would take until 2004 for the first experimental forecasts to be made using convection allowing models (CAMs; Hazardous Weather Testbed 2021). In the following years, operational CAMs were developed and deployed (e.g. the High Resolution Rapid Refresh (HRRR) model; Benjamin et al. 2016). Soon after, NCEP created the High Resolution Ensemble Forecast system (HREF) that combined the various deterministic operational CAMs into an operational convection allowing ensemble (CAE; Roberts et al. 2019).

However, despite these advancements in NWP for forecasting severe weather, key limitations remained. First, even the most frequently updating CAM was limited to hourly data assimilation (Benjamin et al. 2016). While CAMs are able to produce realistic storm evolutions, this lack of frequent data assimilation precludes them from producing a one-to-one correspondence with observed storms (Stensrud et al. 2009). Secondly, CAMs produce deterministic model output that does not reflect the uncertainties with rapid storm evolution that is highly sensitive to the surrounding environment and internal processes (Stensrud et al. 2009). While the HREF ensemble alleviates this problem to an extent and allows for probabilistic forecasting, the individual CAMs within the HREF all suffer from the inability to simulate observed storms described above (Stensrud et al. 2009).

To address these issues, Stensrud et al. (2009) envisioned a NWP system with rapid data assimilation that would allow forecasters to warn on the forecast rather than detection of a storm. Stensrud et al. (2009) suggested that, to achieve this goal, a high resolution (i.e. storm-scale resolution), rapidly updating, convection allowing NWP ensemble that included data assimilation of radar and satellite data to allow for initialization and explicitly prediction of the evolution of existing convection would be needed (Stensrud et al. 2009).

The National Severe Storms Laboratory (NSSL) began work on the Warn-on-Forecast System (WoFS) in 2009 (Heinselman et al. 2023). WoFS is a high-resolution (3 km grid spacing) CAE with 18 forecast members that assimilates radar and satellite data every 15 minutes to improve the modeled representation of storm scale features in real time (Heinselman et al. 2023). While innovations in NWP and the real-time assimilation of radar and satellite data significantly help to improve storm-scale forecasts (Heinselman et al. 2023), WoFS can still struggle with storm location and evolution like other CAMs, particularly early in a storm's lifecycle (Guerra et al. 2022). Additionally, while WoFS updates more frequently than other CAMs with initializations every 30 minutes, the time between initializations can make it challenging to capture rapidly intensifying/developing severe weather events (Heinselman et al. 2023). Nevertheless, the development of WoFS represents a major milestone in watch-to-warning forecasting as the first operational model focused entirely on the 0-6 hour forecast period (Heinselman et al. 2023).

1.2.2 CAE Severe Hazard Probabilities

Even if WoFS or other similarly high-resolution NWP models were able to perfectly place and develop storms, hail cores, regions of severe wind, and especially tornadoes often have too small of spatial and temporal scales to be explicitly

represented with current NWP techniques. To account for this, various explicity modeled NWP fields (updraft helicity (UH), updraft speed, low-level wind speed, etc.) have been used as proxy predictions for the various severe hazards (Kain et al. 2008, 2010; Sobash et al. 2011; Naylor et al. 2012; Sobash et al. 2016, 2019). These fields, when used in an ensemble such as the HREF, can be used to produce probabilistic forecasts of individual severe hazards. Further, these predictions have proved skillful from the day 1/2 outlook timescale (Sobash et al. 2016) down to rolling 4-hour forecasts (Jirak et al. 2020).

1.2.3 Machine Learning (ML) Forecast Guidance

Machine learning (ML) techniques have been applied to many forecasting challenges in the atmospheric sciences beyond severe weather (Monteleoni et al. 2013; Williams 2013; McGovern et al. 2015; Wang et al. 2016; Wimmers et al. 2019). Among the strengths of AI tools are the ability to handle large amounts of data and merge multiple datasets into one cohesive message (McGovern et al. 2023; Loken et al. 2022b, in review). Further, while the computational cost of training ML models can be relatively high (particularly for deep learning models), once training is complete, they can produce forecasts using orders of magnitude less time and energy than traditional NWP models (Bauer 2024).

ML applications to severe hazard forecasting include products intended to provide guidance at multi-day (mimicking SPC convective outlooks; Loken et al. 2020; Hempel 2022; Hill et al. 2023a), watch (intended to provide "first guess" watch fields; Harrison et al. 2022), watch-to-warning (smaller spatial scale or storm-based probabilities for the 0-4 hour period; Flora et al. 2021; Loken et al. 2022b, in review), and now-casting lead times (storm-based probabilities of imminent severe weather; Cintineo et al. 2020; Sandmael et al. 2023; Cintineo et al. 2024; Schmidt et al. 2024).

While non-ML techniques had already been applied to the longer timescales of severe hazard forecasting described above, ML techniques often proved to outperform non-ML baselines due to the ability of ML tools to use more NWP forecast fields and learn relationships between them (Loken et al. 2020; Clark and Loken 2022; Harrison et al. 2022).

Both WoFS ML Severe (Flora et al. 2021) and WoFS-PHI (Loken et al. 2022b, in review) forecast individual severe hazards in the watch-to-warning space and employ similar tree-based ML methods. However, among the key differences is that WoFS ML Severe uses only WoFS NWP data, but WoFS-PHI also uses now-casting information from ProbSevere version 2 (PS2; Cintineo et al. 2020) to improve upon near-term probability magnitude and location. Also important is that WoFS ML Severe operates in an object-based framework. This means that it can only produce probabilities where WoFS produces a storm, but it allows for focused probabilities. WoFS-PHI (Loken et al. 2022b, in review) produces spatial forecasts. This allows for forecasts to be non-zero regardless of the presence of a storm produced by WoFS but can lead to diffuse probability fields at times.

1.2.4 Challenges with ML Guidance

One nearly universal weakness across most machine learning methods for explicit severe hazard forecasting discussed (including those outside of the watch-to-warning space) is that they are trained to predict local storm reports (LSRs; Cintineo et al. 2020; Loken et al. 2020; Flora et al. 2021; Hempel 2022; Harrison et al. 2022; Sandmael et al. 2023; Cintineo et al. 2024; Loken et al. 2022b, in review). LSRs have long been used as a verification metric for warnings and severe convective outlooks (National Weather Service 1982, 1986), so this has made them a straightforward solution to the question of how to define severe "events" on which to train machine

learning models. However, there are known biases in geographic and temporal reporting of severe weather. There is a bias towards a higher density of LSRs near urban areas (Edwards et al. 2018), LSRs are more common during daylight hours (Allen and Tippet 2015), and there are more severe wind reports in the southeastern US than the central US despite the central US being more susceptible to severe winds due to a higher density of damage indicators (trees, powerlines, etc.) in the eastern US (Smith et al. 2013; Edwards et al. 2018). This can lead to ML models trained on these data learning to reproduce these nonphysical biases in their forecasts.

In an attempt to address these biases, Flora et al. (2025) explored using radar-derived Multi-Radar Multi-Sensor maximum estimated size of hail (MESH) as a target dataset rather than hail LSRs for WoFS ML Severe hail forecasts and found that it improved forecast quality. As such, finding alternative or additional sources of "truth" data that better represent severe hazard coverage could similarly improve WoFS-PHI forecasts. Part of this work will examine the potential for using hazard specific severe warnings (in addition to LSRs) that could better represent severe coverage and improve the quality of WoFS-PHI watch-to-warning forecasts.

1.2.5 Initial Version of WoFS-PHI

This project will focus on further development of WoFS-PHI. WoFS-PHI combines PS2 (Cintineo et al. 2020) with WoFS data (Heinselman et al. 2023) using a random forest (RF) classifier to produce hazard specific spatial probabilities for 30 minute time periods with lead times of up to 3 hours. One of the novel aspects of WoFS-PHI is its combination of observation-based data (PS2) with NWP data (WoFS) to produce forecasts. It should be noted that while PS2 probabilities are themselves produced by a machine learning, these models were trained on radar, satellite, and lightning data (with some NWP environmental fields) (Cintineo et al.

2020), and PS2 data can be conceptualized as pseudo-observational data in interpretations of WoFS-PHI. While the combination of observational and NWP data using machine learning is not necessarily a new concept (Williams et al. 2008), its application to severe hazard forecasting is. Williams et al. (2008) used a random forest to combine radar/satellite and NWP predictors to produce thunderstorm occurrence forecasts for aviation and found that both the observational and NWP data were important for different forecast situations. Given these results, it is unsurprising that WoFS-PHI produced higher quality forecasts than machine learning models trained on only WoFS or only PS2. It was also found that observation-based data contributed most to forecasts at early lead times, and WoFS data contributed most to forecasts at later lead times (Loken et al. 2022b, in review).

WoFS-PHI was trained to predict hail, wind, and tornado LSRs for 30-minute forecasts extending out to three hours (Loken et al. 2022b, in review). WoFS-PHI was first tested in the 2023 Spring Forecasting Experiment (Clark et al. 2023) and Watch to Warning Experiment (Berry et al. 2023). Objective (Loken et al. 2022b, in review) and subjective (Clark et al. 2023; Berry et al. 2023) feedback suggested that WoFS-PHI was reliable and able to highlight regions of severe weather effectively. Further, the utilization of PS2 predictors led to improved forecasts compared to a baseline model trained only using WoFS predictors, especially at early lead times (Loken et al. 2022b, in review).

However, while WoFS-PHI was reliable, its forecasts were not particularly sharp, and it struggled to generate high-end probabilities (approximately $\geq 60\%$ for wind and hail and $\geq 30\%$ for tornadoes; Loken et al. 2022b, in review). This was noted by forecasters in both experiments that suggested that the low probabilities made it difficult to determine when there was a higher threat level, and the diffuse nature of the probability fields created challenges in identifying corridors with locally higher

threat levels (Clark et al. 2023; Berry et al. 2023). This study will seek to address these issues by using improved (PS3) and additional (TORP) predictor data. Further, including warnings in the target dataset will be tested in order to identify the predictor/target datasets that lead to the highest quality forecasts.

1.3 Research Questions and Hypotheses

Experiments to test potential strategies to improve WoFS-PHI watch-to-warning forecasts can be divided into two main groups. One set of experiments will focus on changing the predictors, and the other set of experiments will focus on changing the target dataset.

The research questions associated with the first set of experiments are:

Q1a: Does adding TORP predictors improve the skill of WoFS-PHI
Q1b: Does using PS3 predictors instead of PS2 predictors increase forecast quality?
Q1c: How does the combination of PS3 and TORP predictors influence the quality of WoFS-PHI forecasts?
Q1d: If neither PS2 nor PS3 are used, will TORP predictors be a viable substitute for ProbSevere predictors?

The hypothesis associated with these experiments is as follows:

H1: Both TORP and PS3 will improve forecasts for all hazards, but they will be most impactful for tornado forecasts. Hail and wind forecasts will benefit more from PS3 data than from TORP data.

Since tornado forecasts have the least skill in the original WoFS-PHI model (Loken et al. 2022b, in review), these predictions are expected to see the biggest increase. TORP is expected to influence tornado forecasts more than hail and wind forecasts since TORP is specifically designed for tornado forecasting. However, PS3 does have specific hail and wind aspects, so PS3 is expected to have more of an influence on hail and wind forecasts than TORP. To test H1, a baseline version of WoFS-PHI (using WoFS and PS2 without TORP - the same configuration as in Loken et al. (2022b, in review)) will be compared to a version that uses WoFS, PS2, and TORP to isolate the impact of TORP. Then, another version of WoFS-PHI will be trained using WoFS, PS3, and TORP to isolate the impact of PS3.

The research questions associated with the second set of experiments are:

Q2: Will including hazard-specific warnings in the target dataset improve the skill of WoFS-PHI?

The hypotheses for these experiments are:

H2: The change to the target dataset will increase the climatology of events in the target dataset. As such, forecast sharpness and quality will improve due to the RF having more instances of events to learn from. In addition, it is expected that including warnings will help to correct the underrepresentation biases associated with LSRs, and this will also serve to improve forecast quality.

A version of WoFS-PHI that uses WoFS, PS3, and TORP will be trained against LSRs and warnings. This will then be compared to the LSR-only version of the same

model trained to answers Q1b and Q1c. The same metrics for evaluation and explainability to compare the models will be used as for Q1.

1.4 Thesis Organization

Chapter 2 of this thesis will begin with a detailed discussion of the data used for predictors and labels in the new versions of WoFS-PHI as well as how the models were developed, verified, and analyzed for explainability. Chapters 3 and 4 will focus on verification and explainability of the different WoFS-PHI models. Chapter 5 will apply the results that were found to two case studies of example forecasts. Lastly, Chapter 6 will provide a discussion and review of the research questions to synthesize all results that were found and what the results indicate for necessary future work.

Chapter 2

Data & Methods

2.1 Data

2.1.1 WoFS Training Features

WoFS is an 18-member CAE that implements a novel approach of rapid assimilation of radar and satellite data to improve severe weather forecasting in the watch-to-warning period as detailed above (Heinselman et al. 2023). WoFS provides the NWP component of the predictors for WoFS-PHI with both environmental and storm-attribute variables used to extract predictors. One notable limitation of WoFS is its limited spatiotemporal domain. Due to the high computational expense associated with running WoFS and its experimental nature (during the development of WoFS-PHI), WoFS runs were limited to 900 km x 900 km grids centered on region of expected high-impact weather for a given day (Heinselman et al. 2023). WoFS has 3 km grid spacing, which leads to a 300 x 300 forecast grid for a total of 90000 grid points.

2.1.2 ProbSevere Training Features

ProbSevere is an object-based machine learning algorithm designed to identify and track storms and determine the likelihood that they produce any severe hazard within the next 60 minutes (Cintineo et al. 2020). ProbSevere creates objects by

finding regions of ≥ 40 dBZ (or local maxima within a region of 40+ dBZ) which gives storm objects a spatial component beyond a simple point based procedure (Cintineo et al. 2020). PS2 is the currently operational version of ProbSevere; however, version 3 (PS3) has recently been developed, and it delivers substantially improved forecasts of all three hazards (Cintineo et al. 2024). Key differences that allow for these improved forecasts are the switch from a naïve Bayesian classifier to gradient boosted trees, the switch from lower resolution Rapid Refresh NWP data to higher-resolution HRRR NWP data, and the inclusion of additional radar and satellite data in PS3 compared to PS2 (Cintineo et al. 2024). While PS3 produces forecasts of greater quality than PS2, both deliver the same forecast fields with probability of severe hail (ProbHail), probability of severe wind (ProbWind), probability of tornadoes (ProbTor), and probability of any severe hazard.

2.1.3 TORP Training Features

TORP is an operational tornado detection tool that aims to build on previous algorithms such as the Tornado Detection Algorithm (TDA; Mitchell et al. 1998). TORP is a random forest-based tool that utilizes radar data to make probabilistic predictions for whether a tornado *currently* exists at a point (Sandmael et al. 2023). It should be noted that while there is some evidence that TORP probabilities can provide insights regarding tornadogenesis and tornado strength, it is only trained to predict tornado existence.

Similar to ProbSevere, TORP gives point-based probabilities (as opposed to spatial probabilities). However, while ProbSevere objects identify entire storms (and thus have an inherent spatial component that can be used to map objects to a grid), TORP identifies points with no spatial dimensions. Another important difference between ProbSevere and TORP is that, as noted above, TORP is trained only as a

detection algorithm, whereas ProbSevere produces probabilities of a storm being severe in the following 60 minutes. Thus, while TORP has been shown to outperform other diagnostic tornado algorithms and provide some prognostic capabilities (Sandmael et al. 2023), the extent to which it would help to improve a prognostic model is unclear.

2.1.4 WoFS Preprocessing

There are 54 WoFS fields that are used as predictors in WoFS-PHI (Loken et al. in review). For each variable, the temporal maximum or minimum (or both, depending on the variable) value during each forecast period was identified. Then, the ensemble mean of the temporal maximum/minimum values were used (Loken et al. 2022b, in review). In the case of $2-5$ km updraft helicity (UH), both the ensemble mean and the individual member values were retained due to their known importance for WoFS severe weather forecasting (Clark and Loken 2022).

RFs forecast for each grid point independently. Thus, any spatial relationships must be passed explicitly and intentionally as predictors to an RF. To account for this, spatial maximum and minimum values of predictors (referred to as "convolutions" henceforth) are used with radii of 15 km, 30 km, 45 km, and 60 km. It should be noted that Loken et al. (in review) uses square radii, but circular radii are used in this project to obtain more symmetric fields. For WoFS predictors, these convolutions are only applied to storm attribute variables with the exception being that convolutions of minimum sea level pressure and minimum surface pressure (environmental variables) are used, and no convolutions of maximum predicted hail size from Hailcast (a storm attribute variable) are used (Loken et al. 2022b, in review). Generally, the decision on whether to take a maximum or minimum convolution was done to match the temporal maximum/minimum for each variable,

but a complete list of WoFS variables and convolutions used can be found in Table 1 of Loken et al. (in review) along with further details on WoFS preprocessing. There are 174 WoFS predictors after preprocessing is complete.

2.1.5 ProbSevere Preprocessing

Just 18 ProbSevere variables are used to generate 90 corresponding predictors in WoFS-PHI. All ProbSevere objects have an associated storm motion, and this motion is used to extrapolate the objects to create a hypothetical swath representing the path the object would take if its storm motion was maintained up to and through each forecast period (Loken et al. 2022b, in review). Figure 2.1 shows an example of how an object is extrapolated to create a swath. The object probability, spatially smoothed object probabilities, 14 and 30 minute changes in object probability (raw probability), object age, and extrapolation lead time were used. The object age was allowed to be as high as 3 hours if an object with the same ID was persistent for 3 hours preceding the initialization time. There was one of each predictor for each hazard for a total of 18 ProbSevere fields, and each field utilized the same spatial convolution approach as described with WoFS for a total of 90 predictors.

All points on the native WoFS grid that overlapped each swath were assigned the same original ProbSevere object attributes except for the smoothed probabilities and extrapolation lead time. The smoothed probabilities were obtained by smoothing the final field of raw probabilities for each forecast period, and the extrapolation lead time represents the number of minutes that the original object was extrapolated forward to arrive at that gridpoint (Loken et al. 2022b, in review). In the event that no ProbSevere object or convolution overlapped with the WoFS grid, default values

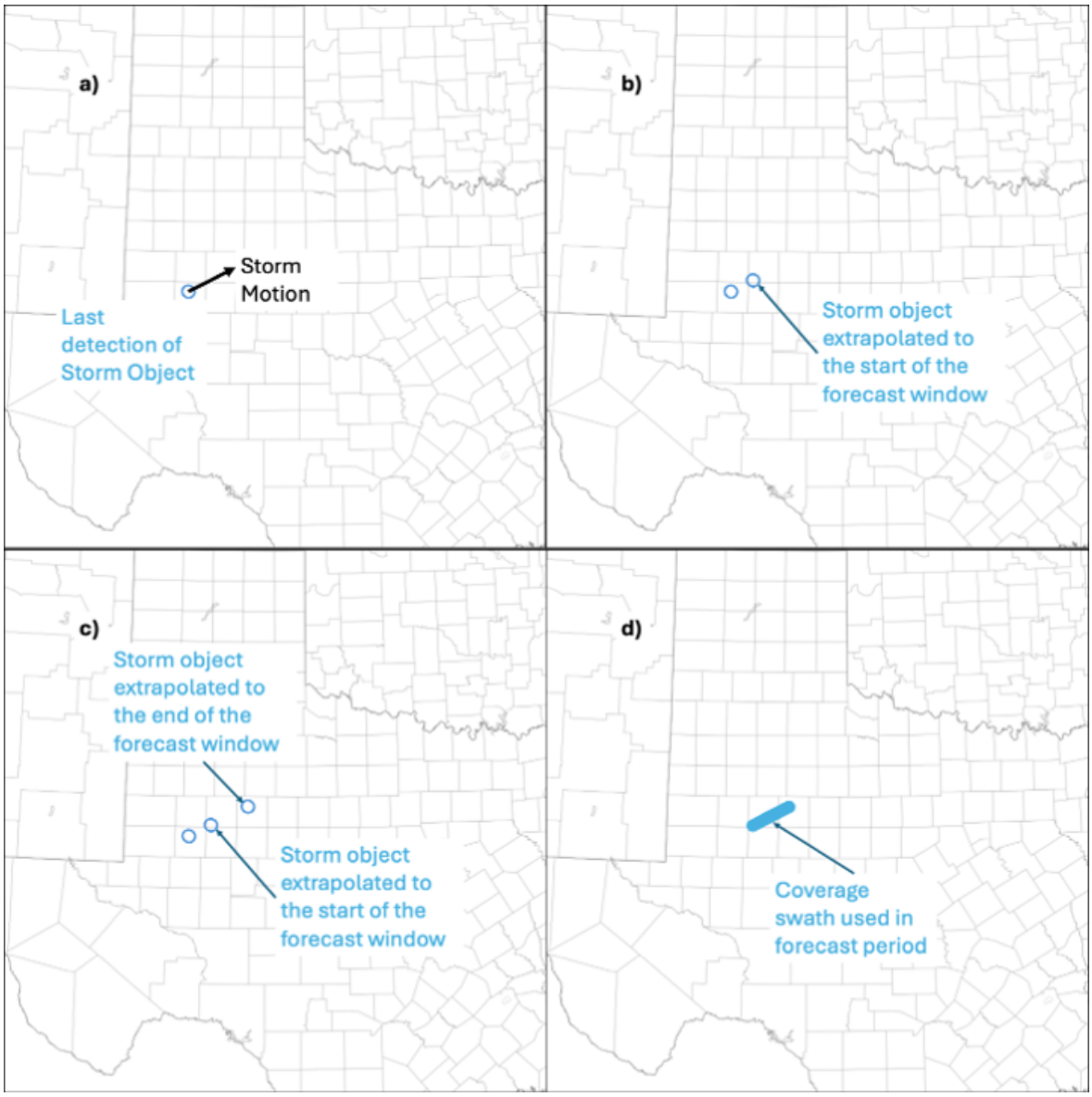

Figure 2.1: An example of mapping a storm object to the WoFS grid (a), extrapolating it to the start of the forecast period (b), extending the extrapolation to the end of the forecast period (c), and determining the final storm swath (d) for an theoretical forecast period.

of 0 were assigned for raw and smooth probability and probability changes, and a -1 was used for object age and extrapolation lead time.

Because the same forecast fields are produced in PS2 and PS3, the same preprocessing method can be applied to the new PS3 data. Further details regarding the preprocessing of ProbSevere data can be found in Loken et al. (2022b, in review).

2.1.6 TORP Preprocessing

As noted, TORP identifies points as opposed to spatial objects (like ProbSevere objects). To account for this, a uniform, circular $7.5km$ buffer is applied to each TORP point (Gesell 2020) to provide it with a spatial component (thus creating a "TORP object") prior to mapping it to the native WoFS forecast grid. TORP objects identified between 185 minutes and 5 minutes before the first window of a forecast period are considered for utilization in that forecast. All objects associated with the same storm during this time are treated as one object. However, in order to qualify for inclusion in a forecast, the object must have existed between 35 and 5 minutes before the first forecast period. While ProbSevere files are produced in regular 2-minute intervals, TORP objects are identified on irregular radar scanning intervals. As such, a range of time is searched for a recent TORP object, whereas for ProbSevere, the most recent file can conveniently be used to check for current ProbSevere objects. An example timeline of WoFS, ProbSevere, and TORP preprocessing is shown in Fig. 2.2.

Similar to ProbSevere, the TORP algorithm calculates storm motions associated with each object. When including a TORP object in the forecast, the most recent TORP object associated with the storm is extrapolated according to its storm motion out to the start of the forecast period. From here, the position of the TORP object at the start and end of each possible forecast period is also determined from the object's storm motion, and a $7.5km$ buffer is applied surrounding the swath between the start

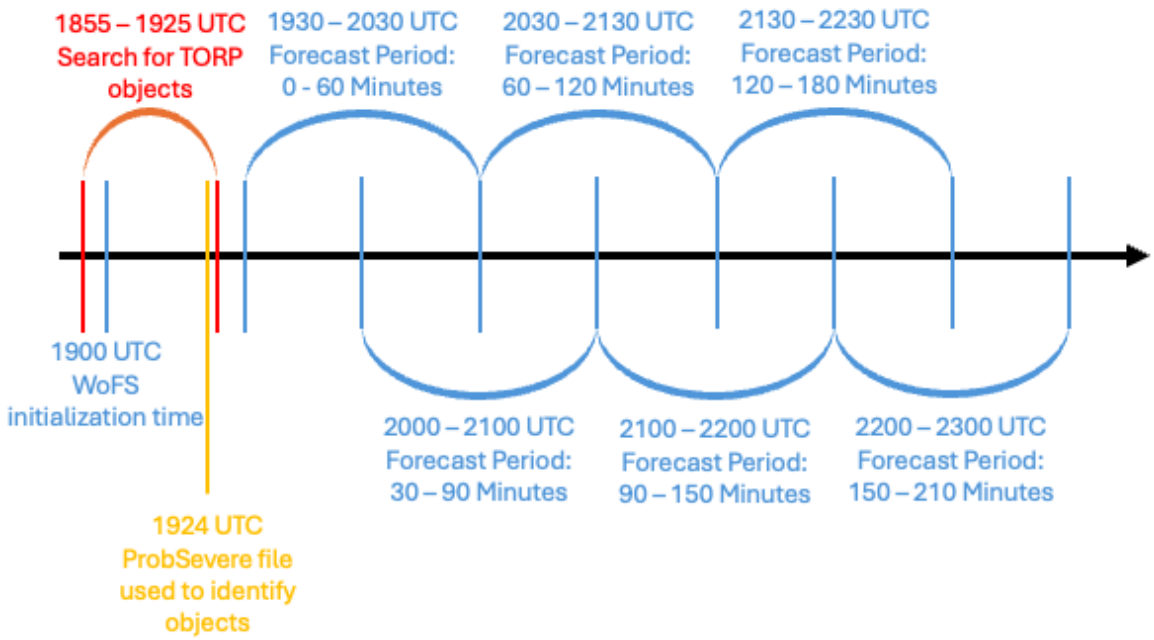

Figure 2.2: An example timeline using a 1900 UTC WoFS initialization. The WoFS, ProbSevere, and TORP initializations and lead times/forecast periods are shown.

and end. An example of a 1-hour forecast of extrapolated ProbSevere and TORP objects is shown in Fig. 2.3.

There are 62 radar-based features that are used to assign a probability to a TORP object (Sandmael et al. 2023), but given that already very large number of predictors in WoFS-PHI prior to the inclusion of TORP (269 predictors) and the fact that only some predictors would be available in real-time, only the 14 potentially real-time predictors were used in WoFS-PHI. In addition to these 14 radar-based predictors, the TORP probability associated with the object, its age, its change in probability (5 and 10 minutes), and the extrapolation lead time of the object were used. The derived TORP predictors (age capped at 180 minutes, probability change, and extrapolation lead time) were implemented to their counterparts that appear in ProbSevere data to make the data sources as similar as possible. However, TORP objects often have relatively short lifetimes (and tornadoes have generally shorter

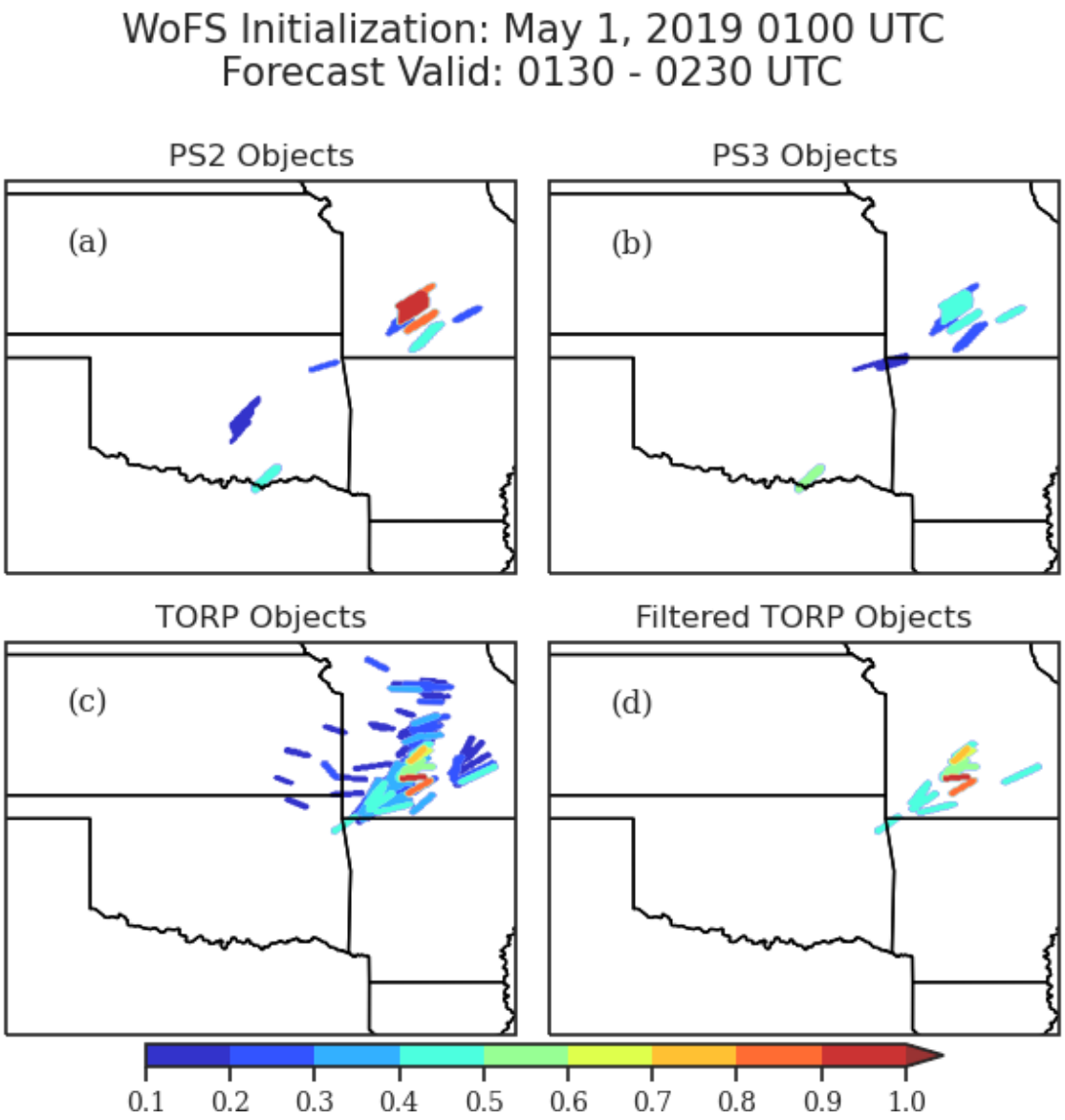

Figure 2.3: An example of tornado PS (PS2, a; PS3, b), TORP (c), and filtered TORP (d) objects colored by probability. This forecast is from April 30, 2019 using a WoFS initialization from 8:00pm CDT with the forecast valid for 8:30pm-9:30pm CDT.

lifespans than hail and wind), so shorter times were used for the probability change of objects. Also, TORP age was calculated as the time past between the objects detection and its most recent detection. This is done intentionally to avoid any bias towards short-lived objects that were detected close to 30 minutes before the cutoff time since it isn't certain that the object still existed after the most recent detection.

Then, as with ProbSevere, a maximum or minimum convolution of each predictor was taken with radii of 15, 30, 45, and 60 km. After all processing of TORP data, this led to 105 new predictors being added to the predictor dataset. The list of 19 predictors, the convolution type used, and the null value (when no object is located at a grid point) can be found in Table 2.1.

Table 2.1: All TORP predictors used to make predictors

TORP Predictor	Predictor Type	Convolution Type	Null Value
Absolute Velocity Max	TORP Input	Maximum	0
Absolute Velocity Max Trend	TORP Input	Maximum	0
Age	Calculated	Maximum	-1
AzShear Max	TORP Input	Maximum	0
AzShear Max Trend	TORP Input	Maximum	0
Correlation Coefficient Min	TORP Input	Minimum	9999
DivShear Min	TORP Input	Minimum	9999
DivShear Min Trend	TORP Input	Minimum	9999
Probability	TORP Output	Maximum	0
ϕ_{dp} Min	TORP Input	Minimum	9999
Probability Change (5 min.)	Calculated	Maximum	0
Probability Change (10 min.)	Calculated	Maximum	0
Reflectivity Max	TORP Input	Maximum	0
Reflectivity Min	TORP Input	Minimum	9999
Rotational Velocity	TORP Input	Maximum	0
Rotational Velocity Distance	TORP Input	Minimum	9999
Spectrum Width Max	TORP Input	Maximum	0
Spectrum Width Max Trend	TORP Input	Maximum	0
Extrapolation Lead Time	Calculated	Minimum	9999

2.1.7 Aggregation of Predictors

The usage of different datasets is predicated on availability. The biggest limiting factor is the availability of WoFS data. The dataset of WoFS runs used for training WoFS-PHI uses 109 days with the first on April 30, 2019 and the last on June 4, 2021 (Loken et al. 2022b, in review). Due to the experimental nature of WoFS during this time, not every severe outbreak is captured in the training dataset, and some WoFS runs occur on convectively benign days with no observed severe weather. While including these null cases arguably makes the training dataset more robust, there is a noted weakness in a lack of days in the training dataset with numerous tornadoes. However, Clark and Loken (2022) showed that even with just a month's worth of data, a useful model can be trained. So, while the number of days with numerous tornadoes is less than the number of days with numerous hail/wind instances, there are likely still enough tornado days to produce a useful model. However, the point of diminishing returns on how many days for training is ideal is unclear. ProbSevere and TORP were not universally available either, but they were generally (though not always) available for days and times when WoFS was run. WoFS-PHI was trained on any forecast for which all necessary data was available, so this leads to minor differences in the temporal domains of the training datasets. These differences in temporal domain arise in instances where PS2 was available, but one or both of PS3 and TORP were unavailable (more often TORP). TORP was still an experimental product during the training domain, and while it was run for the large majority of days in the dataset, there are some days without TORP data. However, these differences are small enough that they are assumed to have negligible effects.

2.1.8 LSR Preprocessing

The original ground truth on which WoFS-PHI was trained was LSRs obtained from NOAA/National Centers for Environmental Information (NOAA 2024). Similar to TORP objects, LSRs occur at a point in space and have no inherent spatial size. The first step to including LSRs was to map them to the nearest point on the WoFS domain. Gridpoints are binarized points with at least one LSR during the forecast period being assigned a value of 1, and 0 otherwise. Next, to give a spatial element, grid points within some radius of an LSR, as detailed in Loken et al. (2022b, in review), are also assigned a value of 1. The two radii used for LSRs are $15km$ and $39km$. The $39km$ radius matches reasonably well with the forecast resolution of SPC products (daily convective outlooks, watches, etc.; Storm Prediction Center 2017), and the 15km radius matches up more closely with warning level forecasts issued by WFOs. The full details of LSR data and preprocessing methods can be found in Loken et al. (2022b, in review).

It should be noted that this method does not capture the reality of the spatial distribution of severe weather. There is a lack of spatial specificity that results from buffering LSR points with a uniform circular radius that will not represent the true, irregularly shaped, coverage swath impacted by each severe hazard. Additionally, LSRs likely underestimate the coverage and frequency of severe weather across all hazards due to population density, diurnal trends in reporting, and the location of damage indicators to identify severe impacts after an event (Kelly et al. 1978; Trapp et al. 2006; Smith et al. 2013; Blair et al. 2011, 2017; Edwards et al. 2018). Further, a particular problem for severe wind LSRs is a known bias towards the southeastern United States where the ratio of estimated to measured severe winds is higher than in other regions (Smith et al. 2013; Edwards et al. 2018).

2.1.9 Warning Preprocessing

In addition to LSRs, the impact of treating hazard-specific warnings as "truth" is assesssed. While the inclusion of warnings as a ground truth has its flaws in terms of consistency and accuracy of warnings between forecasters, offices, and regions (Klockow-McClain et al. 2021), it is important that the timescale that WoFS-PHI forecasts on makes predicting warnings useful, whereas predicting warning occurrence on a nowcasting timescale would be circular logic. Even when a storm is warned and truly does not produce severe weather (as opposed to unreported severe weather), the decision to issue a warning indicates that the storm likely was either nearly severe or appeared to be approaching severe strength. Regardless, this non-zero chance of severe weather should be captured by a watch-to-warning algorithm that focuses on forecasting threat areas with multiple hours of lead time. Forecasting areas will warnings will be issued is also in the spirit of the WoF paradigm (Stensrud et al. 2009; Heinselman et al. 2023). Thus, the choice was made to test the impact of including warnings in addition to LSRs in the event dataset.

Warnings were obtained from the online archive maintained by Iowa State University (Iowa State Univeristy 2021). All warnings that included a severe wind tag ($\geq$ 60 mph) were used for wind events, all warnings with a severe hail tag ($\geq$ 1 in) were used for hail events, and all severe thunderstorm warnings with a"tornado possible" tag and all tornado warnings (radar indicated or observed) were used for tornado events. The "tornado possible" tags were treated as tornado warnings to increase the number of training points for the model under the assumption that while the storm may not have been tornadic in that moment, it had tornadic potential.

2.1.10 LSRs and Warning Preprocessing

Warnings alone were not used as the target dataset for any WoFS-PHI training described here. Warnings were only included in conjunction with LSRs. In order to create an LSR and warning binarized event grid, the binary grids of LSRs and warnings alone were combined by taking the maximum value at each grid point. Thus, if a grid point was within 15 or 39 km of an LSR or contained within a warning at any time during the forecast period, it was assigned a 1, and it was assigned 0 otherwise.

2.2 Model/Training Details

2.2.1 Model Design

As mentioned above, WoFS-PHI (Loken et al. 2022b, in review) utilizes an RF classifier (Ho 1998; Breiman 2001). RFs are a form of supervised machine learning, so every set of predictors must have an associated label (either an event or non-event). Individual decision trees are created from different subsets of predictors to avoid over fitting, and the mean prediction of the decision trees in the forest is the resulting model prediction for a singular forecast. Forecasts are grid based, thus each grid point is treated as a separate forecast by the model.

WoFS-PHI is trained on 1-hour long forecasts out to 4 hours past WoFS initalization. It should be noted that this is a deviation from Loken et al. (2022b, in review) that used 30 minute forecasts. The longer forecast durations were used to align with activities in the Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE; Clark et al. 2024) and Watch-to-Warning Experiment (Lyza et al. 2024) in May and August/September of 2024, respectively. The WoFS-PHI forecast

with the earliest lead time is valid starting 30 minutes after the relevant WoFS initialization (to account for the real-time lag between WoFS initialization and product availability), and to have 30 minutes between start times. "Lead time" will be used to refer to minutes after the first WoFS-PHI forecast henceforth. A timeline of the forecast periods and lead times relative to the WoFS initialization is shown in Fig. 2.2.

In addition to separate models being trained for each lead time, a different model was also trained for each hazard, forecast length, LSR radius, event type (LSRs-only or LSRs and warnings), and model type (combination of WoFS, PS type, and TORP used, discussed further below). Due to the dimensionality of the number of models to train and experiments to conduct, it was not computationally feasible to conduct hyperparameter tuning. The same hyper parameter values were used as in Loken et al. (2022b, in review) as previous work has found these to be relatively successful for severe weather prediction using RFs (Loken et al. 2020, 2022a; Clark and Loken 2022). These values are shown in Table 2.2.

Parameter	Value
Number of Trees	200
Split Quality Criterion	Entropy
Maximum Tree Depth	15
Minimum samples needed for a node to be a Leaf	20
Maximum features to consider when looking for the best split	Square root of the total number of features

Table 2.2: Hyperparameters of the WoFS-PHI RF that were changed from default along with their values

2.2.2 Domain Details

Most of the 109 WoFS runs used in training of WoFS-PHI occurred over the southern Plains (southern KS through northern TX; Fig. 2.4) with a second, smaller, maximum over the mid-Atlantic (VA and NC). Most WoFS forecasts occurred during HWT experiments, especially the SFE that runs from late April through May. This resulted in 5 runs in April (none earlier than April 27), 42 in May, 21 in June, 23 in July, 12 in August, and 6 in September. Notably, there were no cold season WoFS runs included in training, and this is reflected by the relative lack of WoFS runs from the southeastern US included in the training dataset.

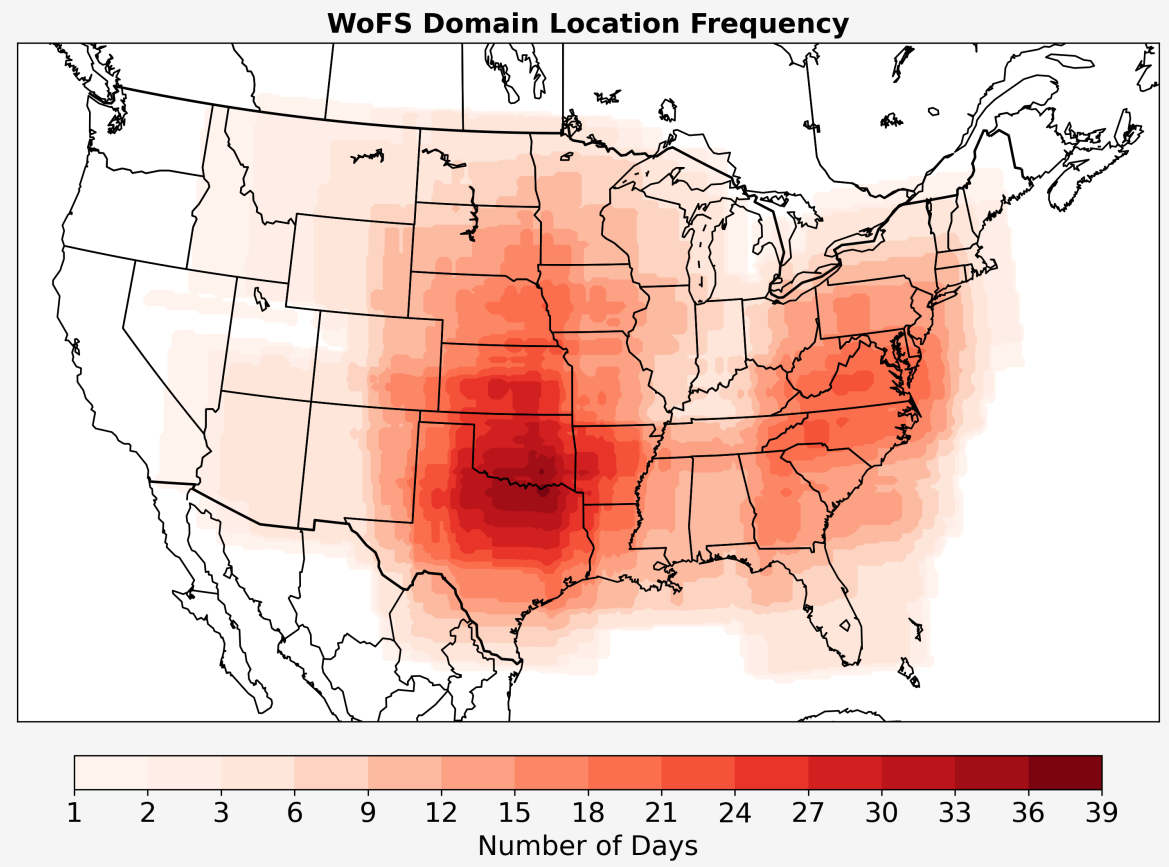

Figure 2.4: A heatmap of WoFS domains used for training WoFS-PHI

2.2.3 K-Fold Cross Validation

Similar to Loken et al. (2022b, in review), k-fold cross-validation was used for training and testing of WoFS-PHI with 5 folds. K-fold cross validation allows for the RF to use all available data for training, validation, and testing. The data is split into k "chunks", where each chunk of data is of approximately equal size. In this work, $n-2$ chunks are used for training, and the $n-1$ and nth chunk are held out of the training process and used for validation (often for hyperparameter tuning) and testing. K-fold cross validation can also be done where $n-1$ chunks are used for training, and the nth chunk is used for validation and hyperparameter tuning. With this strategy, an entirley separate dataset would be used for testing. In this work, after k iterations, each chunk has been held out for testing once. The overall performance of the model can then be evaluated on the aggregated testing data. Here this procedure was used with 5 approximately equally sized chunks with 21 or 22 days per chunk.

Data were split sequentially based on dates to ensure approximate independence between testing, training, and validation sets. It is possible that similar synoptic regimes may have been in place on adjacent days around the cutoff between folds, but it is assumed that this is a minor effect compared to dependencies between data from the same day. As mentioned, hyperparameter tuning was not conducted; however, training, validation, and testing datasets were still used to create a mapping of probabilities to success ratios (SRs) for real-time forecasts. The details of the SR mapping are not discussed here as they behave as a visualization tool for the real-time product that worsen objective forecast quality, and the social science question of whether forecasters prefer this visualization is both yet to be tested in a controlled experiment and beyond the scope of an objective assessment of WoFS-PHI.

Forecasts were trained on a random 10% sample of the available points for training each day to make the training computationally feasible. Doing this reduces the number of training samples in each fold from about 90 million to 9 million and is the same approach taken by Loken et al. (2022b, in review). However, validation and testing were conducted using the full forecast grids.

There are two common strategies for obtaining a real-time model after k-fold cross validation. One is to train the final model on all available data and use that model for real-time forecasting. The second strategy, which is implemented here, is to choose the trained model that produced the best performance on the testing data to be the real-time model (Hill et al. 2023b). The metric used here is the BSS (Wilks 2011), as opposed to the raw brier score (BS) to aid in contextualization of the scores as skill relative to a known climatology.

2.2.4 Experiments Conducted

Experiments were run to investigate the impact of using different training and target datasets. The various model combinations are shown in Table 2.3. It should be noted that all models trained include WoFS predictors. Thus, when discussing models through the rest of the paper, they can be assumed to have WoFS predictors included whether specified or not (for instance, a model referred to as "PS2 No TORP" has WoFS and PS2 predictors, but no TORP predictors).

Model Configuration	WoFS	ProbSevere	TORP	Target Dataset
1	Yes	v2	No	LSRs
2	Yes	v2	Yes	LSRs
3	Yes	v3	Yes	LSRs
4	Yes	v3	Yes	LSRs and Warnings
5	Yes	No	Yes	LSRs and Warnings

Table 2.3: The combinations of predictor and target datasets and forecast durations to be tested in various experiments.

2.3 Evaluation

2.3.1 Verification

Verification is performed in a grid-based manner on the native WoFS-grid. That is, each grid point is treated as its own forecast during verification. Verification is done using three main metrics: the BSS (Wilks 2011), performance diagrams (Hsu and Murphy 1986), and attributes diagrams (Roebber 2009). The BSS is used as an overall evaluation of the model's performance relative to a baseline forecast. In this case, the baseline forecast is a constant forecast of the grid-point-based relative frequency of LSRs (or LSRs and warnings) over the full 109-day dataset. A perfect BSS is a value of 1 which indicates a perfect forecast combined with an imperfect reference forecast (regardless of how imperfect the reference forecast is). A value of 0 indicates that the model performed no better than a reference forecast. The BSS can

be as low as $-\infty$ which represents incorrect forecasts combined with a perfect reference forecast.

Performance diagrams plot probability of detection (POD) against the success ratio (SR) for a set of probability thresholds to create a curve, and goodness is determined by how concentrated the curve is in the top right. POD is defined in terms of the four elements of a continency table where a represents correct "yes" forecasts (hits), b represents incorrect "yes" forecasts (false alarms), c represents incorrect "no" forecasts (missed events), and d represents correct "no" forecasts (correct negatives). Using these representations, POD and SR can be defined as:

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

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

Conceptually, POD represents the percentage of events that were associated with a "yes" forecast, and SR represents the percentage of "yes" forecasts that were associated with an event. In a probabilistic framework, POD and SR can be calculated using various probability thresholds where a probability greater than the threshold is a "yes" forecast and probabilities less than the threshold is a "no" forecast. For hail and wind, thresholds used range from 0% to 100% in increments of 10%. For tornadoes, 0%, 2%, $5\% - 45\%$ (in increments of 5%) and $50 - 100\%$ (in increments of 10%) are used.

One advantage of performance diagrams is that they help to assess forecast quality without giving credit for correct negatives (Roebber 2009). This is especially valuable for rare event forecasting where the vast majority of forecasts are trivial "no" forecasts. Another helpful aspect for severe forecasting is that performance

diagrams show the tradeoff between missed events (POD) and false alarms (SR) at different probability thresholds to assist with decision making since severe hazards may be rare, but they are impactful when they occur.

Attributes diagrams assess the model's reliability and sharpness(Hsu and Murphy 1986). An ideal attributes diagram with perfect reliability will show a straight diagonal line from (0,0) to (1,1). The reliability of the model refers to how well correlated the probabilistic forecasts are to the rate at which events occur. Model sharpness is a qualitative representation of how often a model is producing higher-end forecasts (regardless of how reliable those forecasts are). BSS, performance diagrams, and attributes diagrams are created for each hazard, lead time, time window, and spatial radius by aggregating the forecast data over all relevant testing examples. Thus, only one performance and attributes diagram is created for each model.

A bootstrap resampling procedure is performed to obtain a 95% confidence interval for the BSS values. Additionally, bootstrapping is performed on the BSS calculations by calculating the overall Brier Score (BS) of each WoFS forecast (the full 300x300 grid). The BS values are resampled with replacement n times, where n is the total number of WoFS forecasts for the given configuration, and a mean of the n BS values is taken. This procedure is repeated 10000 times, and the mean BS values at the 2.5th and 97.5th percentiles are noted to generate a 95% confident interval. If the confidence intervals do not overlap, then the difference is statistically significant at the 0.05 level (Knol et al. 2011). However, overlapping confidence intervals are not necessarily reflective of insignificant differences (Knol et al. 2011). In fact, given that the standard deviation of Brier Scores is approximately equal across models and lead times, the probability of non-overlapping 95% confidence intervals is less than 0.01 (Knol et al. 2011).

2.3.2 Explainability

Relative tree interpreter (RTI; Loken et al. 2022b, in review) is used to evaluate the relative importance of predictors to WoFS-PHI forecasts at each point, independent of the RF probability. RTI is used instead of the traditional tree interpreter (TI; Saabas 2016; Loken et al. 2022a) since TI values for a given prediction sum to the probability outputted by the model. When averaged across many forecasts, the "average" contribution by a predictor is consequently disproportionately influenced by the contribution to higher-end forecasts. RTI alleviates this problem by using the relative contribution of each predictor to the prediction, independent of the final RF probability (Loken et al. 2022b, in review). Thus, while TI values sum to the RF probability of a given forecast, RTI values always sum to 1 for any given forecast.

RTI values are calculated for only 10% of the testing dataset using the same method of random sampling as for training due to similar constraints on computational resources. Then, the RTI values are averaged across each forecast in the testing dataset. This allows for an aggregate view of predictor importance across all 5 folds of testing. Additionally, the contribution from like predictors are grouped together by summing their RTI contributions. "Like" predictors are those that are associated with the same physical predictor. For instance, all predictors associated with $2-5$ km updraft helicity are grouped together, and all predictors that are associated with PS tornado probabilities are grouped together. This helps to give a more complete representation of the importance of a predictor that may otherwise be diluted by spreading its importance across many predictors. Further, the aggregate contribution of WoFS, PS, and TORP data were calculated by summing all WoFS, all PS, and all TORP predictor contributions together to evaluate the relative importance of each dataset to the final predictions.

The total contribution by each predictor is also broken down into it's contribution to hits, misses, false alarms, and correct negatives employing the same strategy as in Loken et al. (in review). An important note is that for any given forecast, not all predictors will contribute to the same contingency table category. If an event occurs, predictors that raise the final RF probability are classified as contributing to a hit, and those that lower the probability are classified as contributing to a miss. When an event does not occur, predictors that raise the final RF probability are classified as contributing to a false alarm, and those that lower the probability are classified as contributing to a correct negative.

However, while RTI is useful, it has its drawbacks. It gives disproportionate value to deeper nodes in a tree (Lundberg et al. 2020), and it reduces the contribution by each predictor to a single number, so there is no information regarding how different values of a predictor would influence the final RF probability.

To address these concerns, accumulated local effect was used (ALE; Molnar 2025). ALE provides a more sophisticated approach to answer similar questions as PDP. First, a default prediction, p, is set. Next, the values of the predictor of interest are grouped into bins such that there are an equal number of points in each interval (although the intervals may vary greatly in their ranges of predictor values). The average difference in predictions relative to p is then computed for each instance within each bin to get the local effects. Then, these local effects (LE) are accumulated along the successive bins such that the ALE value in bin n (ALE_n) is: $ALE_n = \sum_{i=1}^{n} LE_i$. Lastly, the ALE curve is recentered such that the mean effect is 0. This re-centering around 0 allows for the ALE curve to represent the average difference from the average prediction for a given predictor at different values (Molnar 2025).

Some weaknesses of ALE include that the binning method can be sensitive to predictors with skewed distributions so that few bins, or occasionally only one bin, can be created. Further, it can only be calculated for individual models and predictors, as opposed to RTI values, which are easily aggregated across models and predictors as described above. To account for this, only the fold with the highest testing BSS is used to compute ALE. This is the same method as is used in determining which model to use in real time for WoFS-PHI. Since predictors that are summed together for an aggregate effect in the RTI analysis cannot be easily aggregated in ALE, the 30 km convolution (when applicable) of each predictor was used since this was the middle point between the value at the point and the 60 km convolution. Also, for the predictors from object based sources (i.e. PS and TORP), the size of the objects in conjunction with the 30 km radius leads to a similar spatial radius as the 39 km used as the buffer surrounding LSRs used to verify WoFS-PHI. It is assumed that while the magnitudes of the impacts may vary across predictors that were grouped together in the RTI analysis, the *pattern* will remain similar across "like" predictors. Further, since each predictor has its own curve, only select predictors could be reasonably analyzed and displayed in the ALE analysis. These predictors are shown in Table 2.4:

WoFS Predictors	**ProbSevere Predictors**	**TORP Predictors**	**Miscellaneous Predictors**
80 m wind speed	Hail PS smooth probability	TORP probability	Latitude
$0-2$ km Updraft Helicity (UH)	Wind PS smooth probability	TORP maximum reflectivity	Longitude
Column maximum vertical velocity	Tornado PS smooth probability		WoFS Initialization Time
$2-5$ km UH			
Supercell Composite Parameter (SCP)			
Significant Tornado Parameter (STP)			
$0-500$ m STP			
Lightning flash extent density			

Table 2.4: The predictors chosen for the ALE analysis organized by dataset type. For each predictor, the 30 km convolution is used (except for the miscellaneous predictors that do not have spatial convolutions).

As with training and the RTI analysis, the same random 10% sample of points in the highest skill fold are used rather than the full dataset due to computational constraints. The exception to this is for TORP probability and TORP maximum

reflectivity. TORP objects represent a skewed predictor that made binning unfeasible. To account for this, all gridpoints associated with a TORP object were used, and 25% of gridpoints not associated with a TORP object were used to compute ALE so that binning could be done more effectively.

Since the same predictors have been chosen for analysis for all hazards, it is expected that some will have minimal impact on predictions for some hazards. Rather than tailor the predictor list to be hazard specific, a uniform list of hazards was kept to show the difference in impact between important and unimportant predictors. This also allows for an easy visual confirmation of one predictor's importance for each hazard to support the premise that WoFS-PHI is learning physically realistic patterns that are specific to each hazard.

Chapter 3

WoFS-PHI Verification

The main form of verification used will be the BSS. However, where BSS confidence intervals do not overlap, attributes and performance diagrams will be shown as well to do additional analysis of the results. While overlapping 95% confidence intervals do not guarantee a lack of statistical significance at the 0.05 confidence level, their overlap (or lack thereof) will be used here to determine which experiments are analyzed in additional depth.

3.1 Impact of Changing the Predictor Dataset

Three experiments focused on changing the predictor dataset are conducted in this section. First, a comparison is done between the baseline model and the baseline model with TORP predictors added (models 1 and 2; Table 2.3) to isolate the impact that TORP predictors have on model performance. Next, a comparison between the WoFS, PS2, and TORP model and a WoFS, PS3, and TORP model is conducted to isolate the impact of PS3 (models 2 and 3; Table 2.3. Lastly, a comparison between the baseline model and the model with both PS3 and TORP (models 1 and 3; Table 2.3) is done to assess the aggregate influence of changing the predictor dataset.

3.1.1 Inclusion of TORP

Including TORP data leads to a small increase in BSS for hail forecasts at early lead times, but there is nearly no change at the later lead times (Fig. 3.1a). However, at all lead times, the confidence intervals overlap. TORP appears to have a similar effect on wind forecasts where the early lead times show higher BSS values when TORP is included. However, the confidence intervals again overlap at all lead times (Fig. 3.1b). Tornado forecasts see the smallest improvement from the inclusion of TORP data, and confidence intervals again overlap at all lead times (Fig. 3.1c). However, the differences in TORP impact between hazards are small since TORP has small impacts for all hazards, and the fact that tornado forecasts see the smallest increase in BSS at the early lead times is likely not significant.

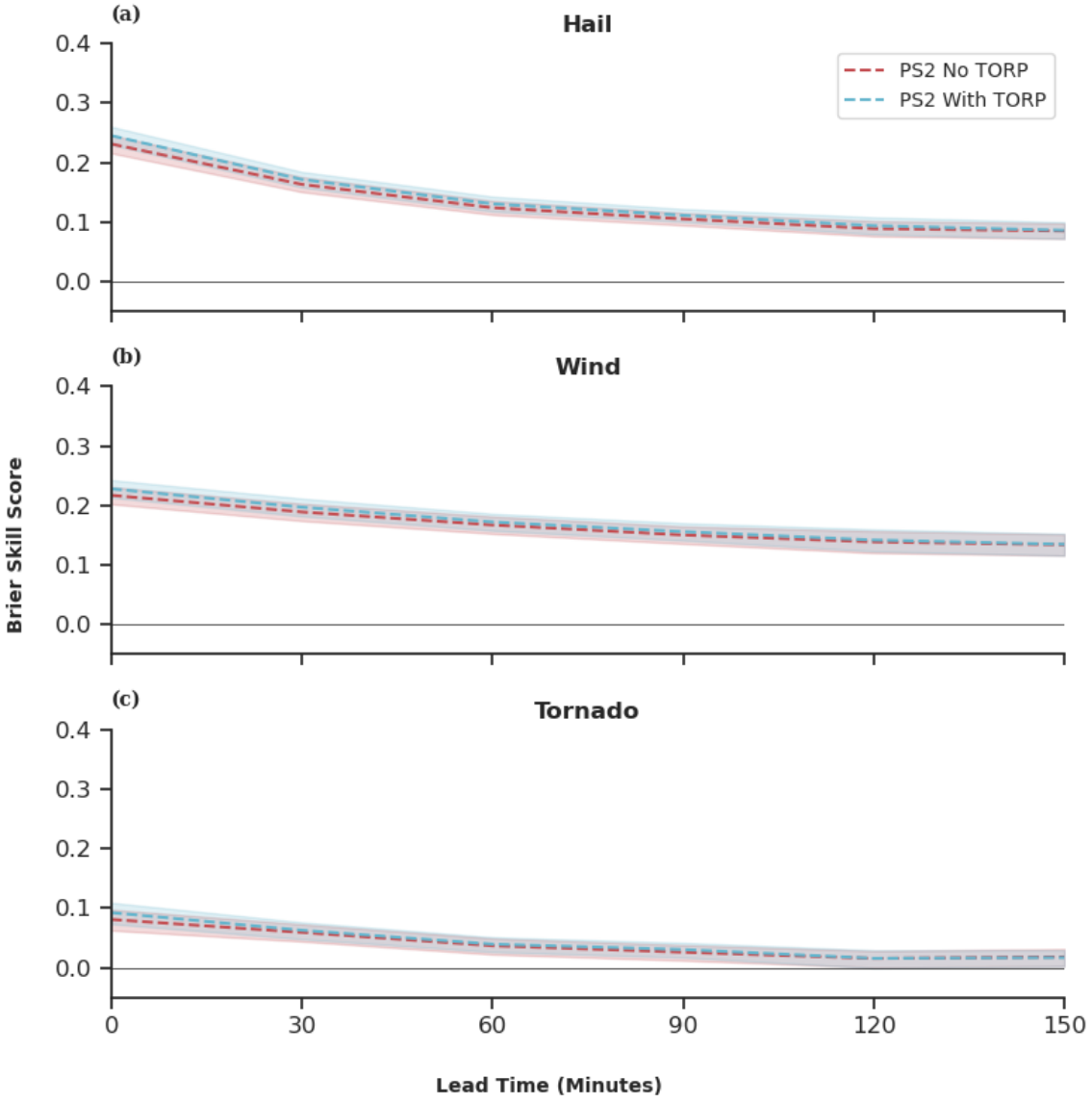

Figure 3.1: The BSS for model configurations 1 (red) and 2 (cyan; Table 2.3) is shown by lead time with the shaded region representing the 95% confidence interval

While the influence of including TORP in predictors appears to be largely marginal, there is some (non-significant) signal that BSS is higher for all hazards (Fig. 3.1) at the earliest lead times. While the small magnitude of influence that TORP had is surprising, the fact that it is focused on the earliest lead times is not. TORP is a tornado detection algorithm based on observed radar data predictors; thus, TORP

probabilities and TORP radar data are designed to be most relevant at the shortest lead times. This finding also makes sense given that Loken et al. (2022b, in review) found the most benefit from including ProbSevere predictors (which are heavily influenced by radar and satellite observations) at lead times of 90 minutes and less.

3.1.2 PS2 vs PS3

The WoFS, PS2, and TORP model from the previous section is used here to compare against a WoFS, PS3, and TORP model to isolate the influence of the ProbSevere version used in the predictors. Using PS3 instead of PS2 predictors leads to an increase in BSS at early lead times. However, similar to including TORP, confidence intervals overalp at all lead times (Fig. 3.2a). The influence of PS3 predictors on wind forecasts smaller than its influence on hail and the influence of TORP on wind forecasts. An increase in BSS can be seen at early lead times, but confidence intervals overlap at all lead times. (Fig. 3.2b). For tornado forecasts, while the BSS confidence intervals still overlap, the overlap is less than what is seen for hail and wind forecasts (Fig. 3.2), especially for the $0-60$ minute forecast period. Forecasts of all hazards saw the greatest improvements to the early lead time forecasts, just as with TORP, and similar to what would be expected based on the nature of ProbSevere and TORP predictors (consisting of or derived from observational data) and the findings of Loken et al. (2022b, in review).

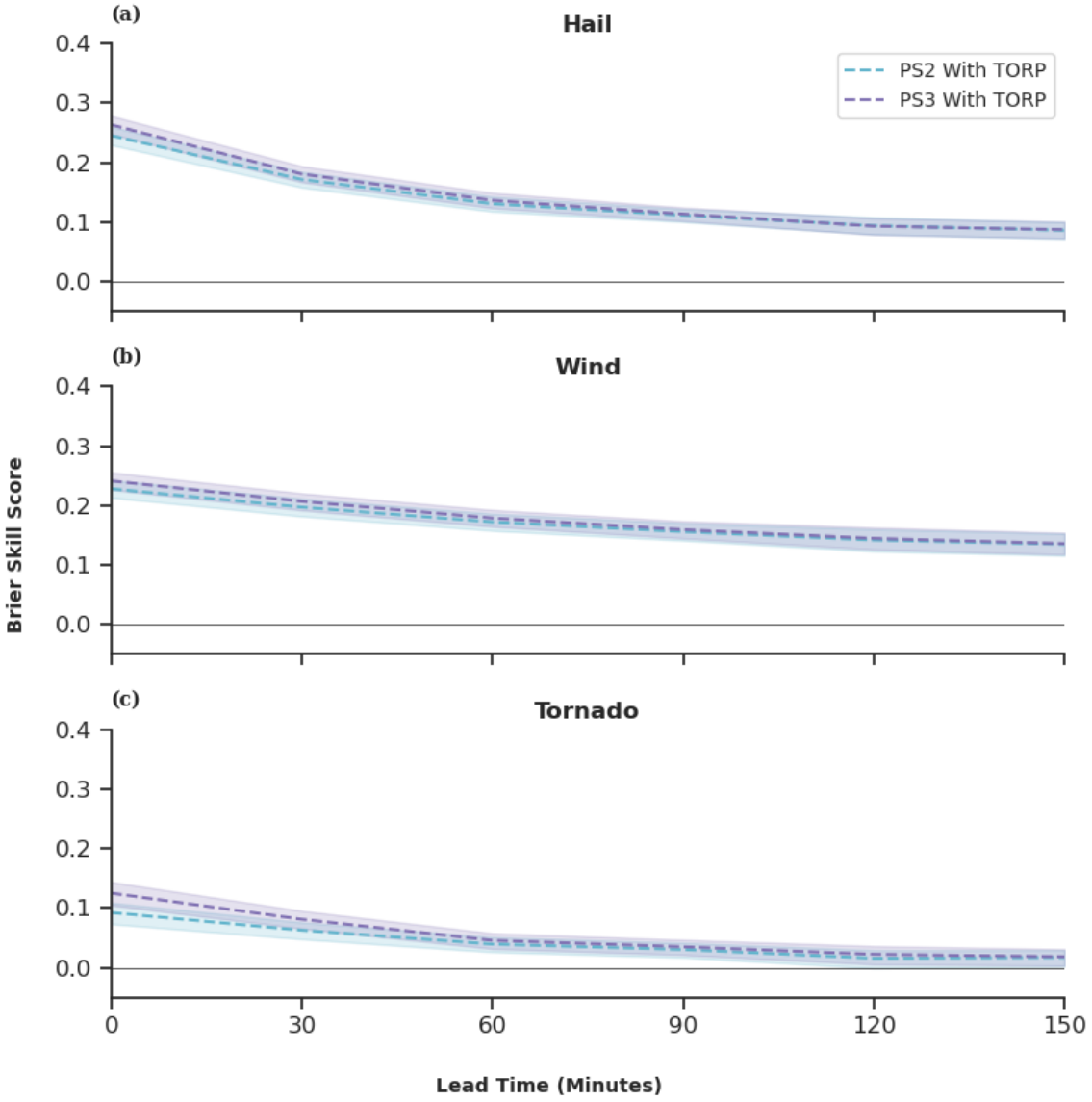

Figure 3.2: As in fig. 3.1 (hail LSRs), except the PS2 with TORP model is in cyan, and PS3 with TORP is in blue.

3.1.3 Impact of Including both TORP and PS3

This experiment consists of comparing the performance of the baseline model (WoFS and PS2) to the WoFS, PS3, and TORP model (models 1 and 3; Table 2.3). While the individual effects of PS3 and TORP were small enough to have confidence

intervals overlap at all lead times for all hazards (except perhaps the impact of PS3 on tornado forecasts), using both PS3 and TORP predictors (as well as predictors from WoFS) had a greater impact on skill. There is a significant increase in BSS for $0-60$ minute hail forecasts with the PS3 and TORP model compared to the baseline and some increase in skill extending to the $90-150$ minute period (Fig. 3.3a). However, confidence intervals overlap for lead times of 30+ minutes. Resolution is also increased for $50-80\%$ probabilities in the $0-60$ minute period, and the peak increase in resolution shifts down to the $40-60\%$ range by the $60-120$ minute period (Fig. 3.4). However, even when assessing the effect of PS3 and TORP together, there is no change to forecast sharpness (Fig. 3.4). The performance diagrams show a rightward shift noticeable out to the $90-150$ minute period, and there is also a more noticeable upward shift associated with an increase in POD for lower probability thresholds (Fig. 3.5).

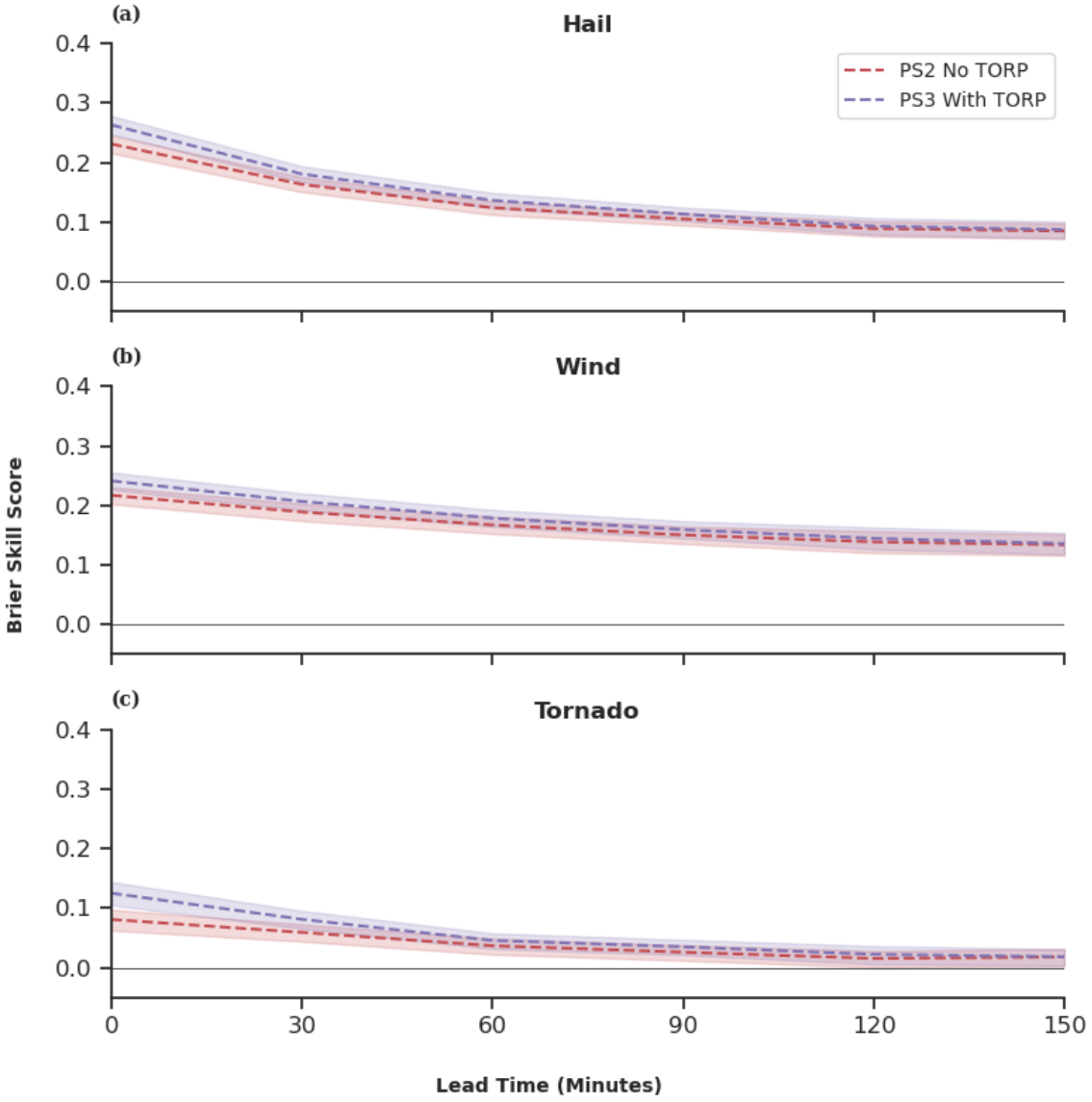

Figure 3.3: As in fig. 3.1 (hail LSRs), except the PS2 no TORP model is in red, and PS3 with TORP is in blue.

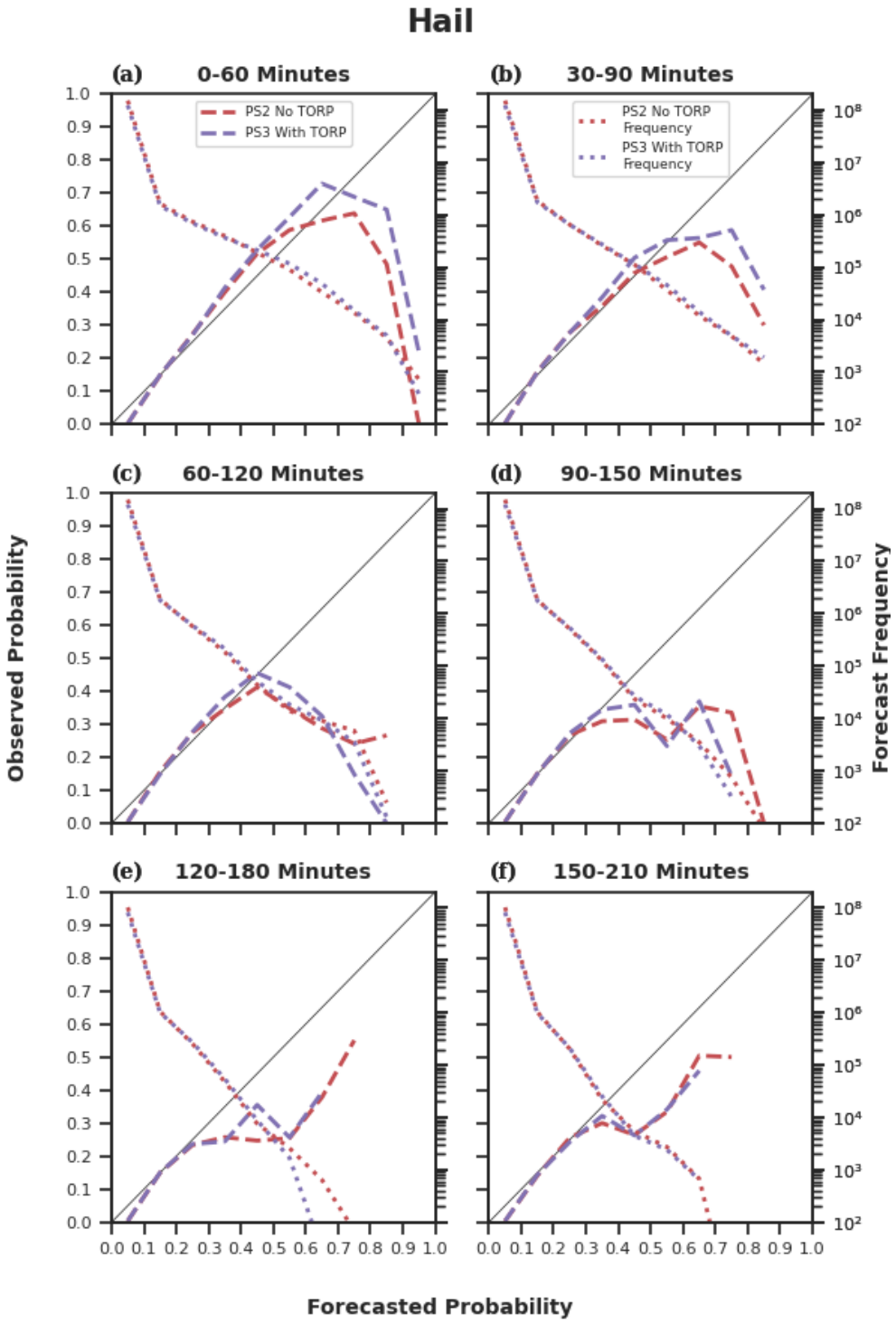

Figure 3.4: Attributes diagrams for the baseline (red) and PS3 with TORP (blue) hail models are shown. The dotted lines represent the frequency of forecast issuance by probability. Panels are alphabetically ordered with lead time (panel (a) is the earliest, 0-60 minute period, and panel (f) is the latest, 150-210 minute period).

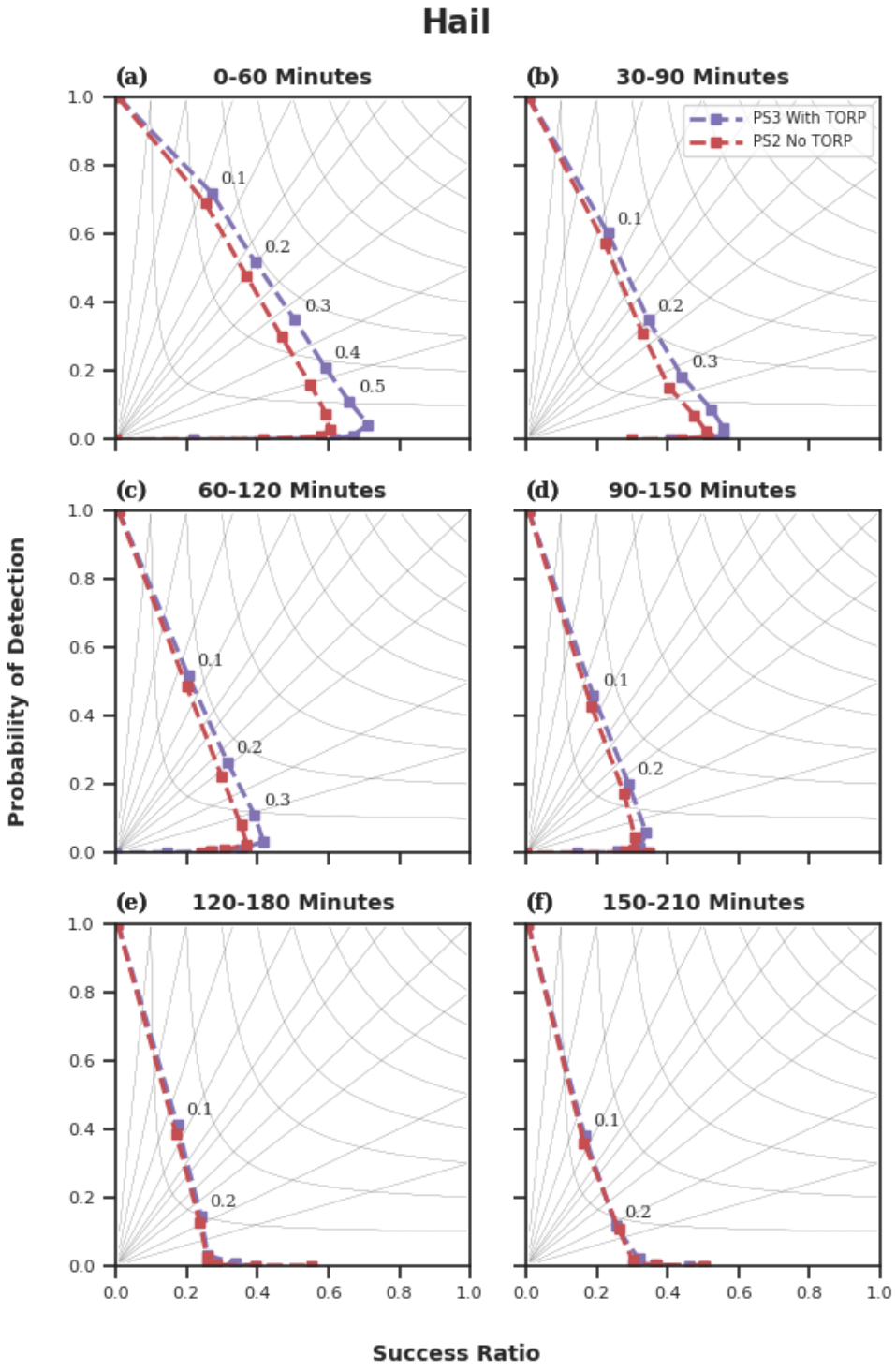

Figure 3.5: Performance diagrams for the baseline (red) and PS3 with TORP (blue) hail models are shown. Panels are ordered by lead time as in Fig. 3.4.

The impact on wind is expectedly the smallest of the three hazards given that both TORP and PS3 individually had the smallest impact on wind forecasts (Figs. 3.1b, 3.2b). However, while the BSS improvement is smallest, and the confidence intervals still overlap even for the $0-60$ minute lead time, the improvement in BSS decreases less rapidly than for hail with some increase in BSS seen out to the $120-180$ minute lead time (Fig. 3.3b).

Given the large improvement from using PS3 predictors instead of PS2, it is unsurprising that tornado forecasts also benefit the most from using both TORP and PS3 predictors. There is significant improvement in the BSS for $0-60$ minute tornado forecasts and some improvement out to the $120-180$ minute period (Fig. 3.3c). There is also an increase in improved resolution and reliability in the $0-60$ and $30-90$ minute forecast periods (Fig. 3.6). There is minimal change to resolution or reliability beyond the 30 minute lead time. However, there is a signal for a *decrease* in sharpness out to 90 minutes of lead time despite the increased BSS, resolution, and reliability (Fig. 3.6). Similar to with hail forecasts, there is an upward shift (associated with increased PODs) that appears for lower probability thresholds and an even larger rightward shift (associated with increased SRs) for higher probability thresholds (Fig. 3.7). There is no rightward shift in the performance diagrams beyond 30 minutes of lead time, but there is still a slight upward shift (associated with increased PODs) found at 60 and 90 minutes of lead time (Fig. 3.7).

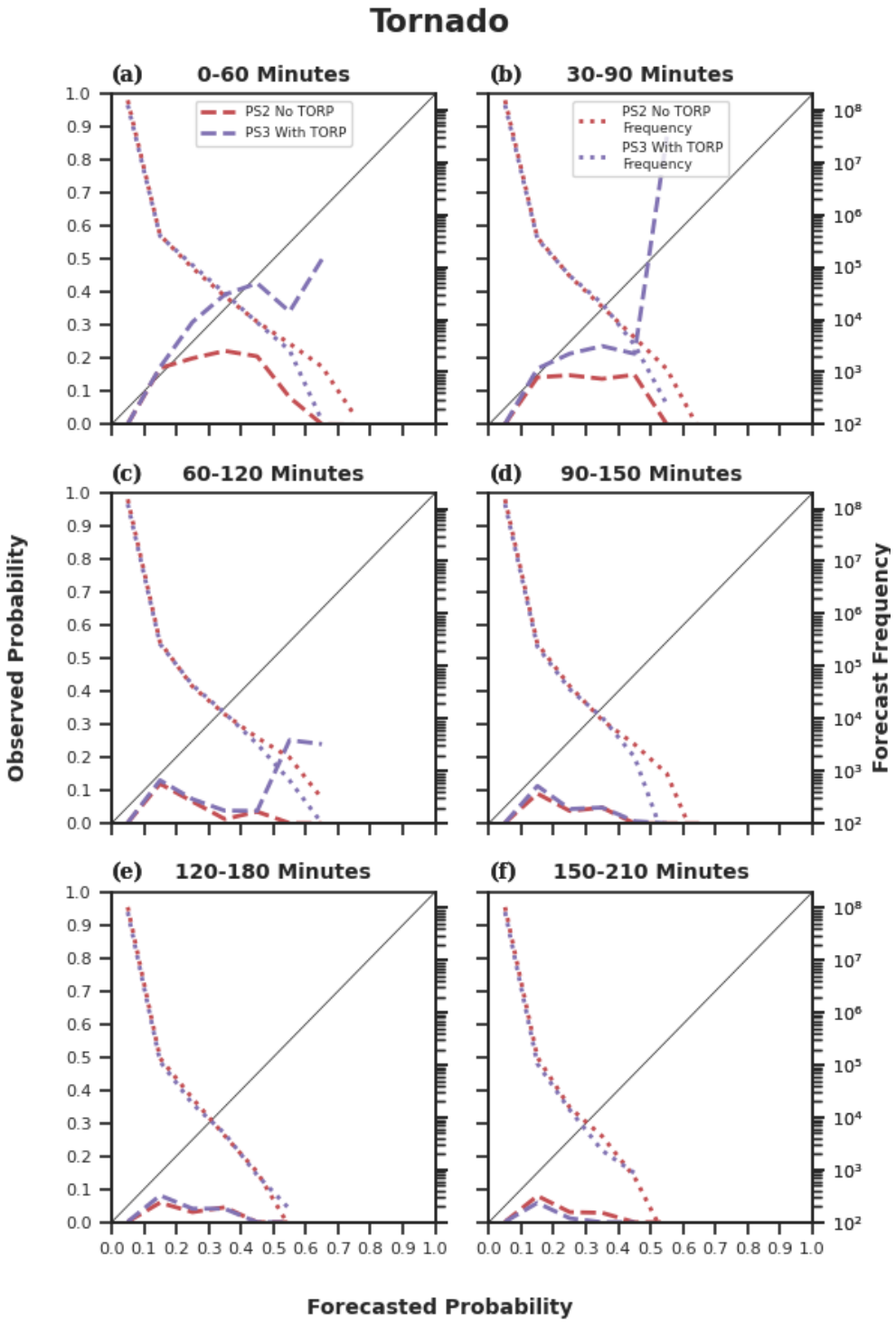

Figure 3.6: As in fig. 3.4 but for tornado LSRs.

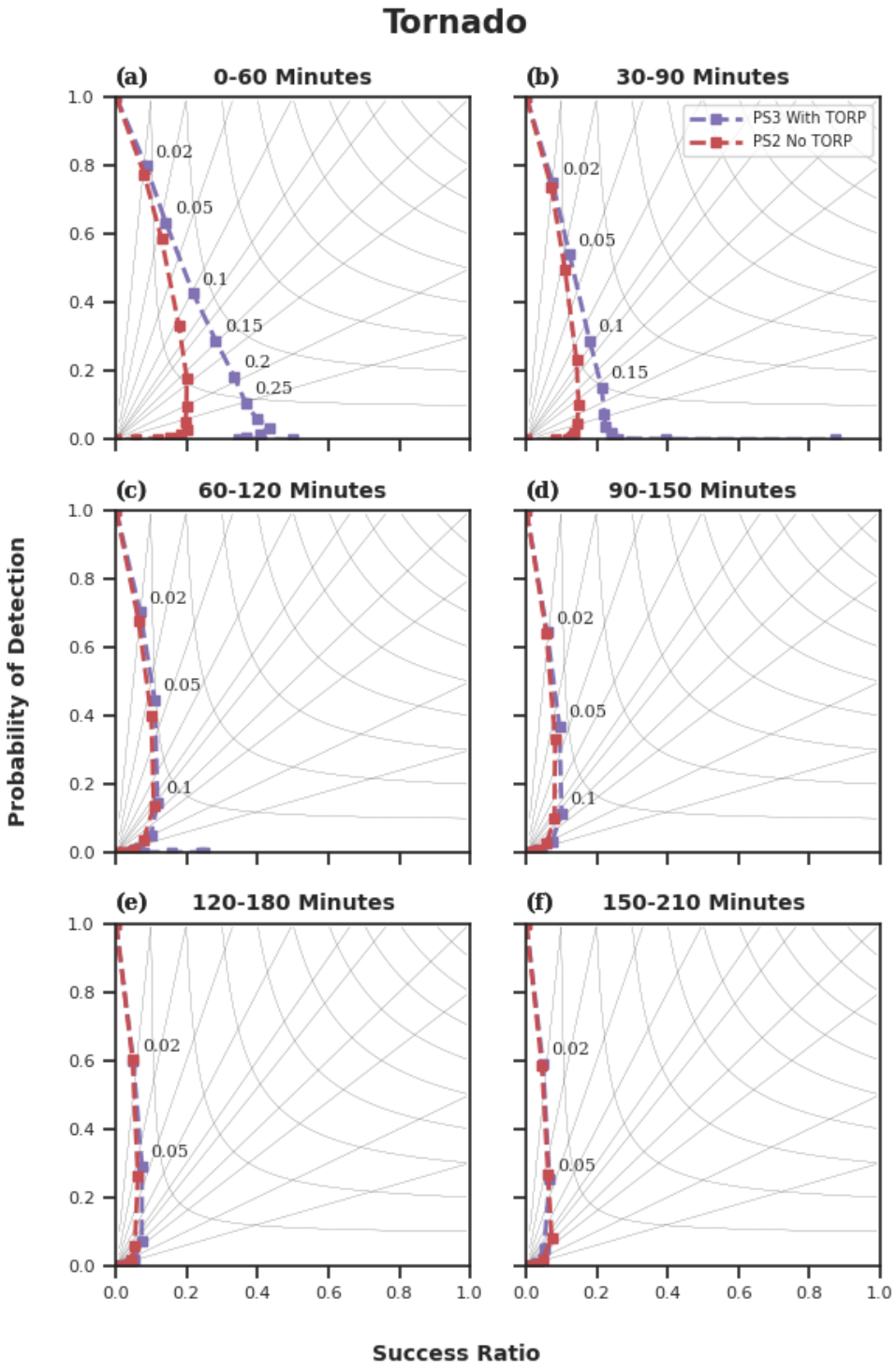

Figure 3.7: As in fig. 3.5 but for tornado LSRs.

3.2 Impact of Including Warnings in the Target Dataset

The next pair of experiments involved testing the impact of changing the target dataset in two different ways. First, hazard specific warnings are used to augment (but not replace) the LSRs in the target dataset for both training and verification (models 3 and 4; Table 2.3). As discussed above, LSRs are known to undersample the occurrence of severe weather due to various spatial and diurnal biases (Trapp et al. 2006; Smith et al. 2013; Allen and Tippet 2015; Edwards et al. 2018). Thus, the inclusion of warnings in the target dataset is intended to help correct the undersampling bias of LSRs. This is evident in the fact that the LSR and warning target dataset has approximately 1.75 (varying between 1.65 and 1.81) times more events than the LSR-only target dataset. However, LSRs still provide value as examples of certain severe weather events (even if the location/timing information can be imprecise), whereas warnings have inherent error as they are human issued forecasts. For these experiments, WoFS, PS3, and TORP predictors were used to train the model as the models trained using these predictor datasets showed often higher (though not always statistically significantly higher), or at least not lower, BSS values relative to the baseline model. To limit the number of experiments done, only one predictor dataset combination was used to test the change to the target dataset, so the PS3 with TORP model was used even if it cannot be definitively said that it is better for all hazards and all lead times compared to the baseline.

There is a significant increase in BSS for hail forecasts at all lead times, and the confidence intervals do not overlap at any lead time (Fig. 3.8a). While this increase in BSS is much larger than what was seen with any changes to the predictor class, it is important to recall that, similar to the experiments between 15 and 39 km LSRs,

the datasets used to verify each model are different. And, when calculating the BSS using a baseline forecast that is the average occurrence of an event, this means that the reference forecast in the BSS is different between the two models shown. So, this significant increase in BSS reflects that the LSR and warnings trained model performs better relative to its appropriate reference climatology than the LSR only trained model.

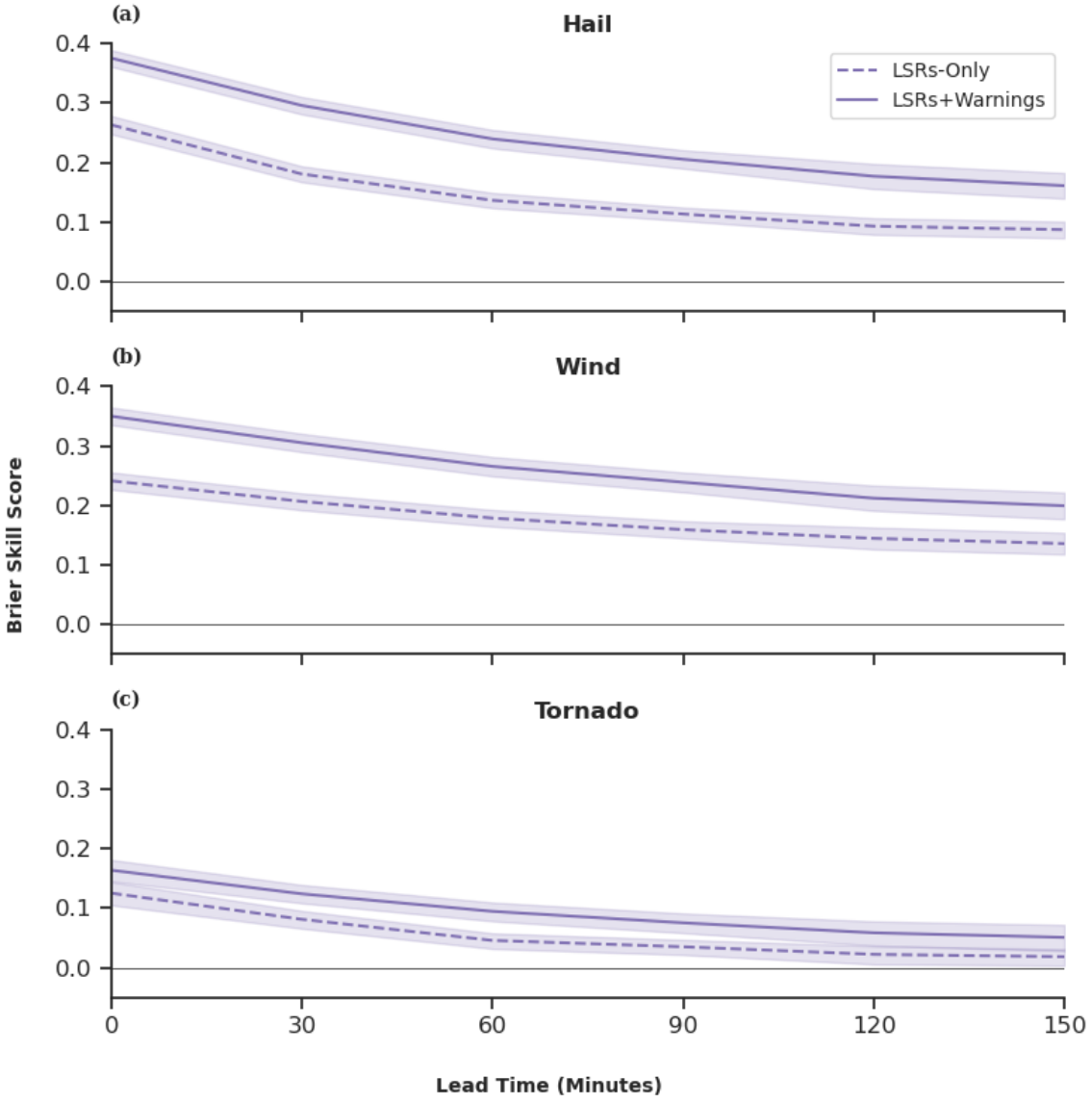

Figure 3.8: As in fig. 3.1 (hail), except both lines represent PS3 with TORP models. The dashed line represents the model trained on LSRs-only, and the solid line represents the model trained on both LSRs and warnings.

For forecasts of $\leq 70\%$ in the $0-60$ minute period, there is a decrease in resolution, but an increase in reliability when training on LSRs and warnings for hail forecasts which indicates a correction of underforecasting (Fig. 3.9). For forecasts of $\geq 70\%$, there is an increase in resolution and reliability. At later lead times, this

threshold shifts to lower values as sharpness decreases with lead time, so the criteria for relatively higher-end forecasts (for which resolution and reliability increase) shifts to lower values (Fig. 3.9). However, the decreased resolution/increased reliability is hardly noticeable beyond the first lead time, and non-existent at 120 and 150 minutes of lead time (Fig. 3.9). Hail forecast sharpness is also higher (something not seen with any changes to the predictor dataset; Fig. 3.9), and this is a metric independent of the verification dataset. Further, the sharpness improvement starts for forecasts that are $\geq 20\%$ at all lead times (Fig. 3.9).

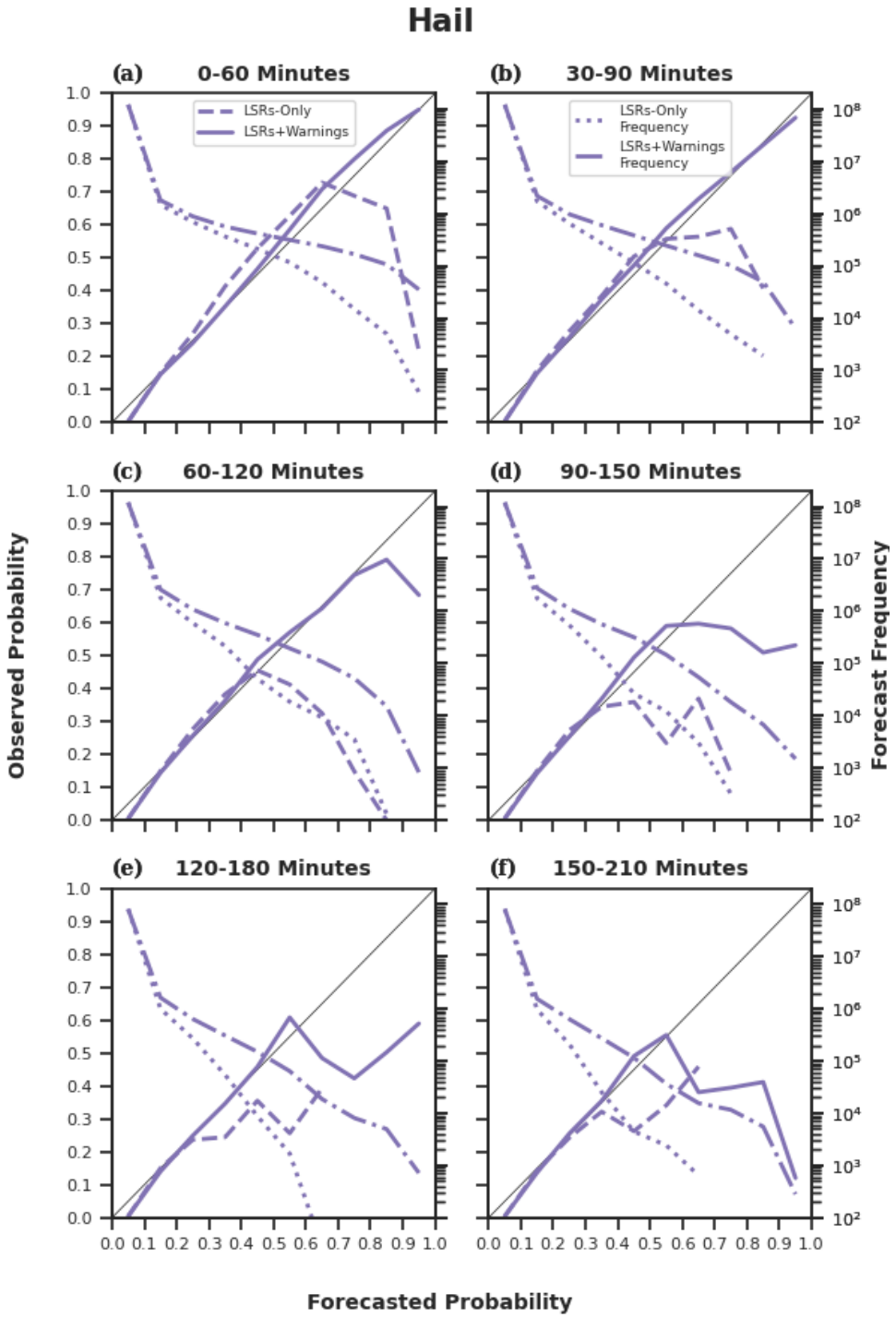

Figure 3.9: As in fig. 3.4 (hail), except both lines represent PS3 with TORP models. The dashed line represents the model trained on LSRs-only, and the solid line represents the model trained on both LSRs and warnings. The dashed-dotted line represents the forecast frequency for the LSR and warning trained model.

Performance diagrams for hail show both a rightward and upward shift at all lead times (Fig. 3.10). Lower probability thresholds see more of an upward shift (increase in POD) than higher thresholds, but higher thresholds see more of a rightward shift (increase in SR) than lower thresholds (Fig. 3.10). This is the same pattern that was seen in the improvement from changing the predictor dataset, though the magnitude of change is much larger with the target dataset change from LSRs-only to LSRs and warnings. Interestingly, while the improvement in SR for higher probability thresholds decreases with lead time, the increase in POD at lower probability thresholds increases with lead time (Fig. 3.10). This is likely linked to the fact that sharpness benefits are seen most at long lead times. Longer lead time forecasts consist of more forecasts at most probability thresholds which indicates that the spatial footprint of the probabilities is larger in addition to the magnitude of the highest probabilities being higher. With a larger spatial footprint, more LSRs/warnings will necessarily be captured, thus increasing the POD. Since the spatial footprint increase is largest for the longer lead times, the POD increase is largest at the longer lead times as well.

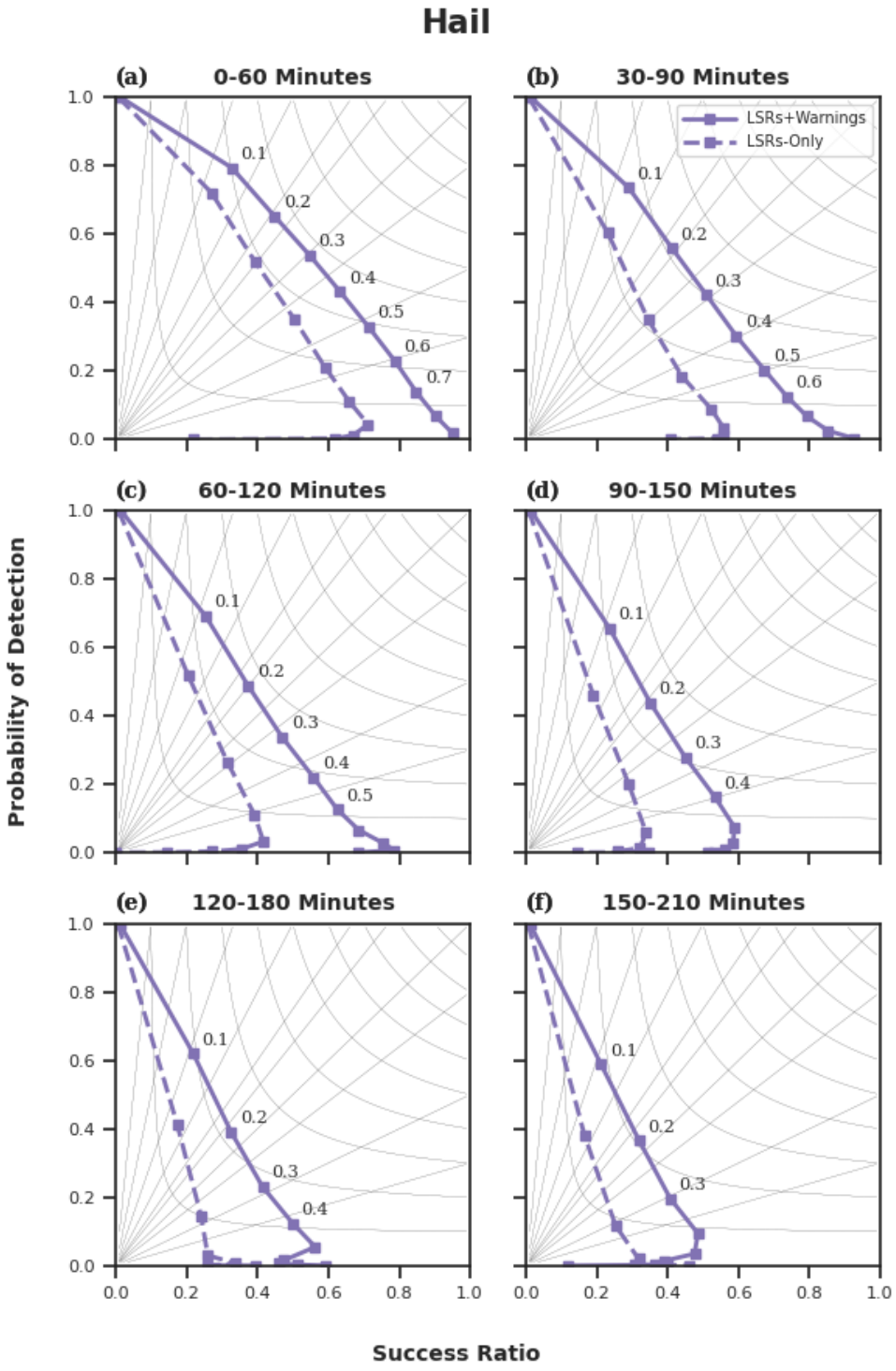

Figure 3.10: As in fig. 3.5 (hail), except both lines represent PS3 with TORP models. The dashed line represents the model trained on LSRs-only, and the solid line represents the model trained on both LSRs and warnings.

Wind forecasts behave similarly to hail forecasts with significant increases in BSS at all lead times (Fig. 3.8b). Also similar to hail forecasts, the improvement decreases with lead time, though it remains a significant improvement at all lead times. In the attributes diagrams, for the first two lead times, resolution is decreased for forecasts $\geq 30\%$ (Fig. 3.11). However, similar to hail forecasts, this improves reliability as the decreased resolution corrects an underforecast. Contrary to hail forecasts, aside from $90-100\%$ forecasts in the $0-60$ minute period, there are no forecasts in the first two lead times that see an increase in both resolution and reliability (indicating that the underforecast was not corrected or made worse; Fig. 3.11). At later lead times, however, forecasts above $50-60\%$ see improvements in resolution and reliability, while forecasts between 30% and $50-60\%$ continue to see decreases to resolution and increases to reliability (Fig. 3.11). In essence, both hail and wind forecasts become nearly perfectly reliable for all forecasts at early lead times and up to 80% (60%) for wind (hail) at later lead times with over and underforecast biases from the LSR only trained model being corrected.

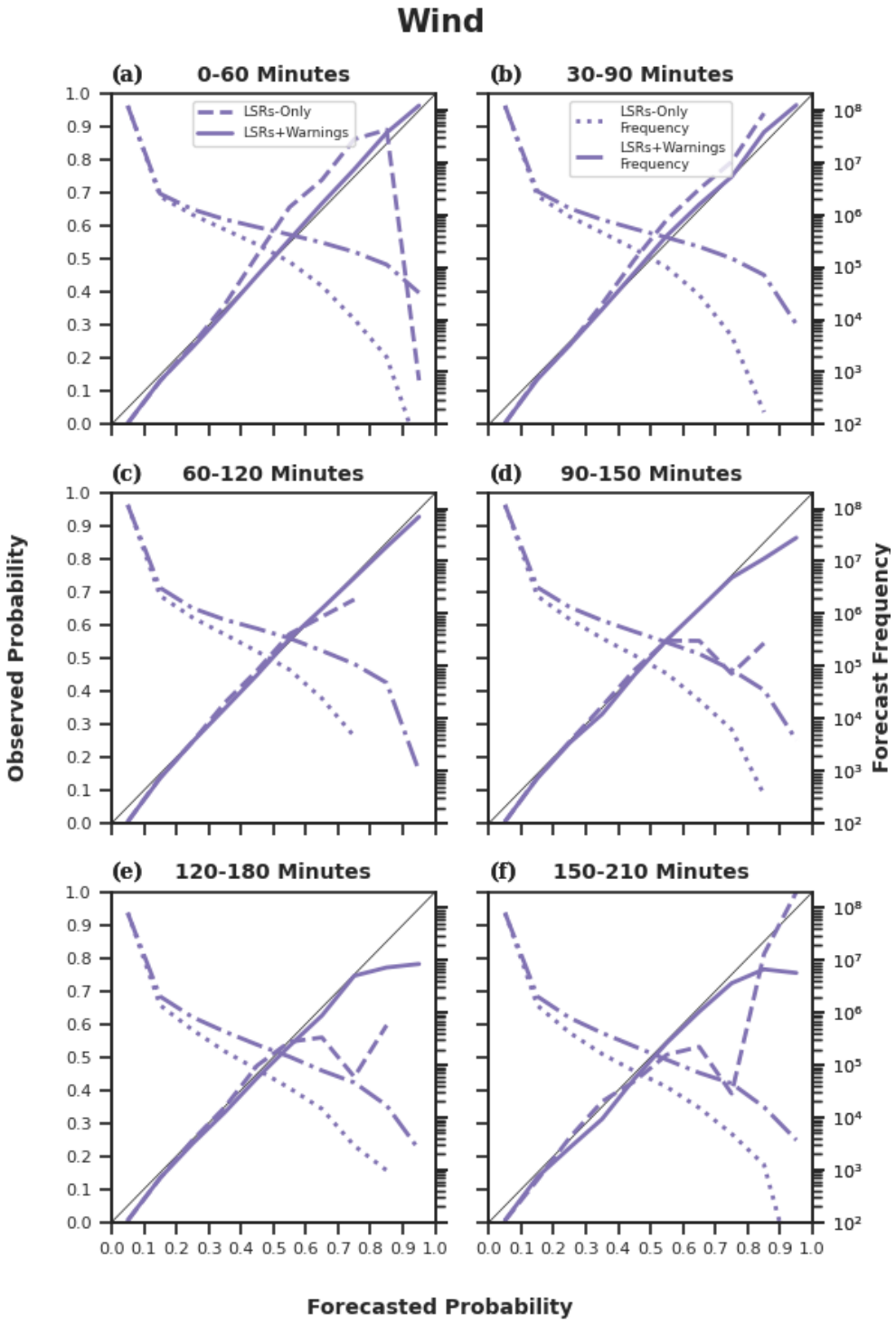

Figure 3.11: As in fig. 3.9 but for wind.

Wind forecast sharpness is also greatly improved, though to a lesser extent than for hail forecasts (Fig. 3.11). Further, the the improvement at later lead times is not nearly as large as it is for hail forecasts. It continues to be shown that wind forecasts are less impacted by lead time than other hazards. This also results in less improvement to POD at later lead times compared to hail, although there is still more of a POD increase at later lead times than earlier ones (Fig. 3.12). However, while the POD for lower probability thresholds does not increase as much for wind as for hail forecasts, the SR of higher end forecasts increases more for wind than hail forecasts (Fig. 3.12). This can be seen by the increased rightward shift of the bottom of the performance diagrams. This is reflected as well in the higher resolution of high-end ($\geq 60\%$) wind forecasts (Fig. 3.11) compared to hail forecasts (Fig. 3.9).

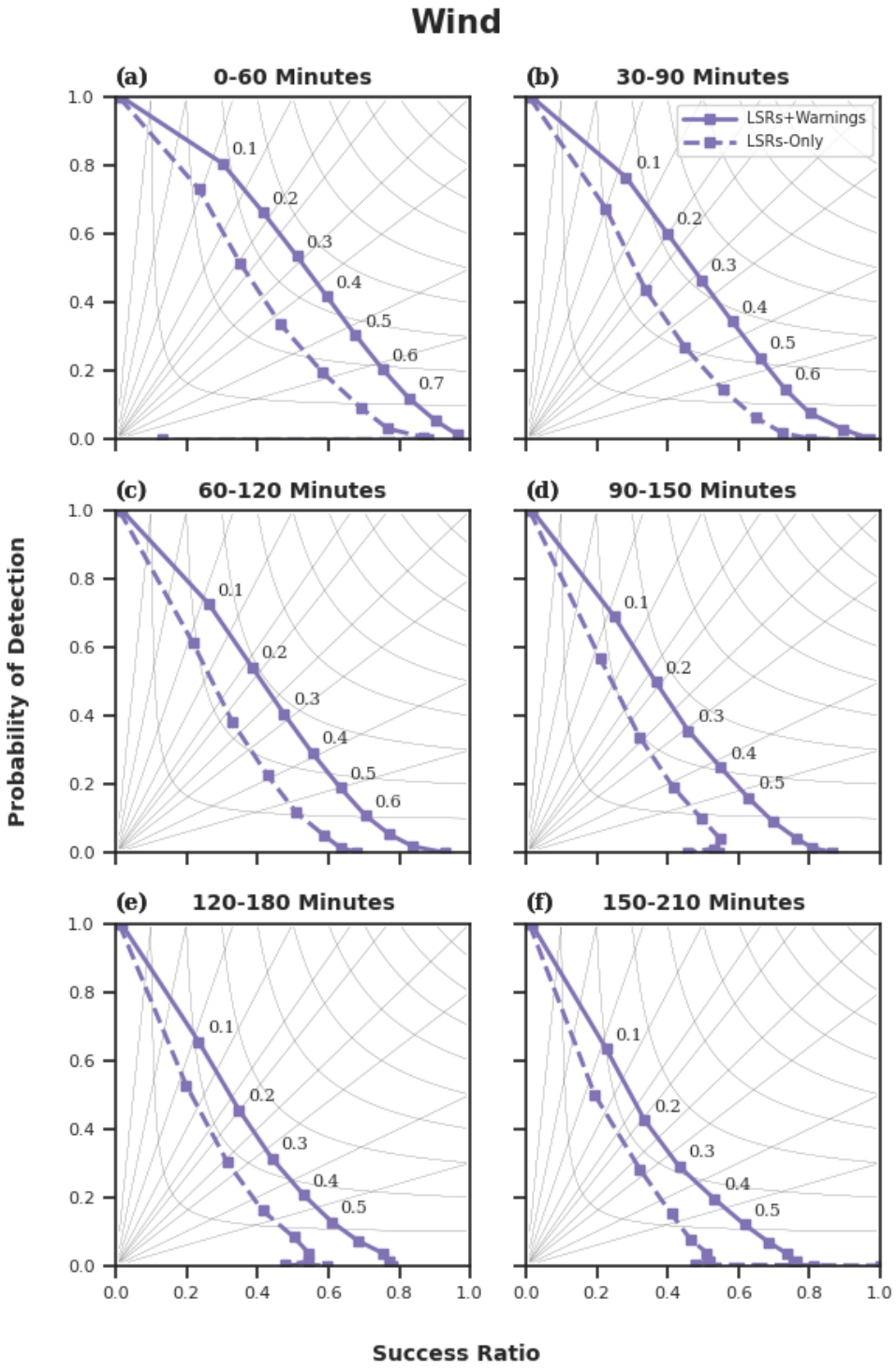

Figure 3.12: As in fig. 3.10 but for wind.

Tornado forecasts behave differently from both hail and wind forecasts. There is still a significant increase in BSS when training on LSRs and warnings, but it is not as large as for hail or wind (Fig. 3.8). Also, while the confidence intervals do not overlap for $0-90$ minutes of lead time, the 120 and 150 minute lead time forecasts do have overlapping confidence intervals, something not seen in hail or wind forecasts (Fig. 3.8). Further, the greatest increase in BSS is not seen at the earliest lead time. Instead, the BSS improvement peaks with 60 minutes of lead time (although the raw BSS value for the LSR and warning model still peaks in the $0-60$ minute and decreases with lead time; Fig. 3.8c). However, despite the fact that the confidence intervals overlap at 120 and 150 minutes of lead time, the BSS increase is still larger than the increase from changing the predictor dataset. Interestingly, the increase in BSS in the $0-60$ minute window is approximately similar to the increase that resulted from changing from PS2 to PS3 predictors, and it is a smaller increase than caused by using both PS3 and TORP predictors.

In contrast to hail and wind forecasts, there are no instances of underforecasting being corrected for tornado forecasts. However, there is still an increase in sharpness for tornado forecasts when trained on LSRs and warnings (as with hail and wind forecasts; Fig. 3.13). There is also an increase in resolution and reliability for higher-end ($40-60\%$ at early lead times, all forecasts at later lead times) tornado forecasts (similar to hail and wind forecasts, though at a lower threshold; Fig. 3.13). Tornado forecasts when trained and verified on LSRs and warnings (LSRs-only) are reliable up to about 50% (50%) at early lead times and to nearly 30% (not reliable at all) at later lead times. Even at early lead times, tornado forecasts between $50-70\%$ are much *more* reliable (even if not perfectly reliable) than the LSR only forecasts in that range (Fig. 3.13).

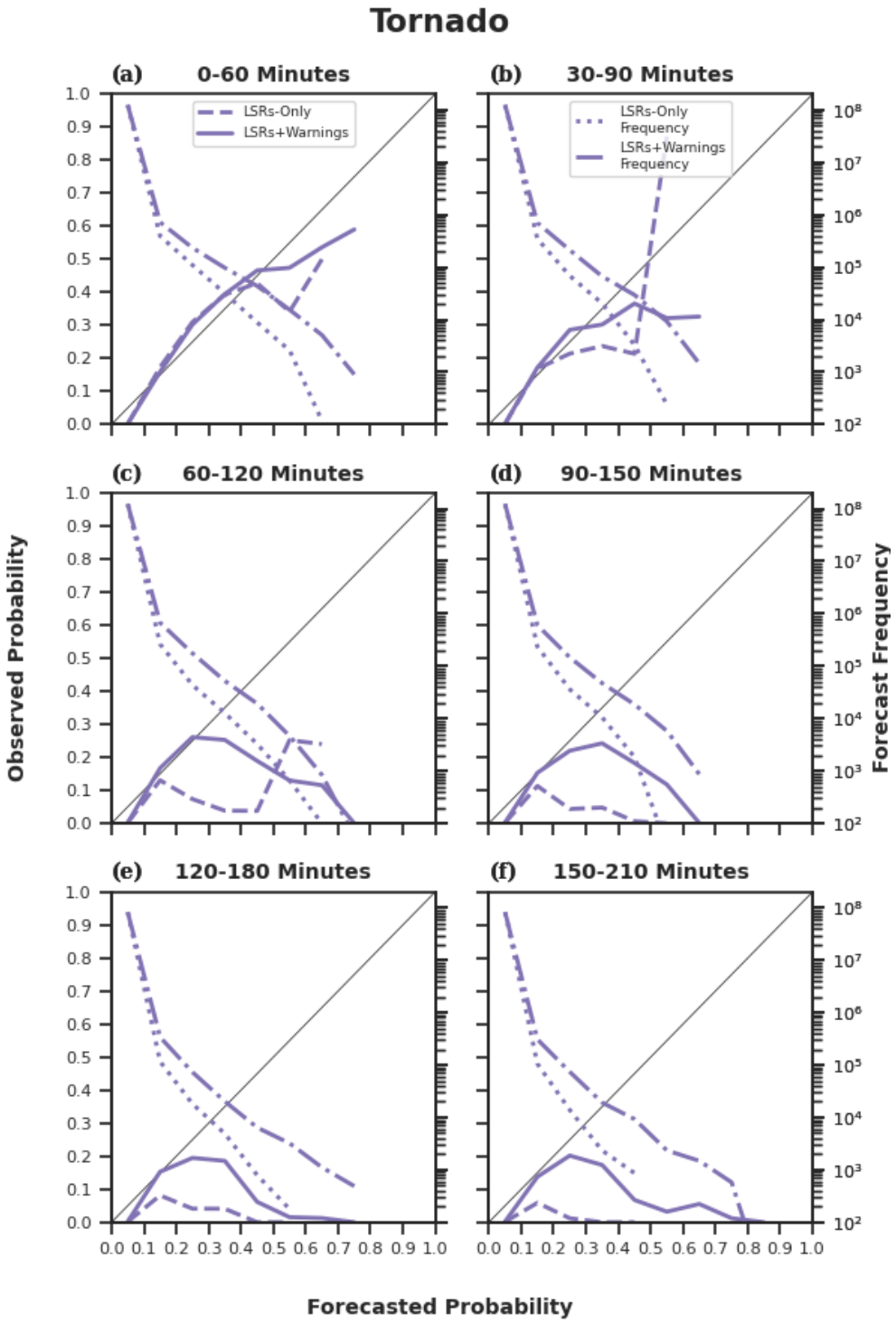

Figure 3.13: As in fig. 3.9 but for tornadoes.

Tornado forecasts see the smallest upward/rightward performance diagram shifts when trained on LSRs and warnings at all lead times (especially early lead times; Fig. 3.14). Consistent with BSS, the biggest increase in the area under the performance diagram curve is with the $60 - 120$ minute window (Fig. 3.14). Further, the increase in POD for lower probabilities at longer lead times is more in line with the large increases seen with hail than the more moderate increases seen with wind. This is consistent with the fact that tornado forecasts saw increases in sharpness that increased with lead time (Fig. 3.13), just like with hail forecasts. SR for higher probabilities increases at all lead times, though they eventually go to 0 for the highest probabilities as reflected by the attributes diagram curve intersecting with the x-axis for lead times of 60 minutes and longer (Figs. 3.13, 3.14).

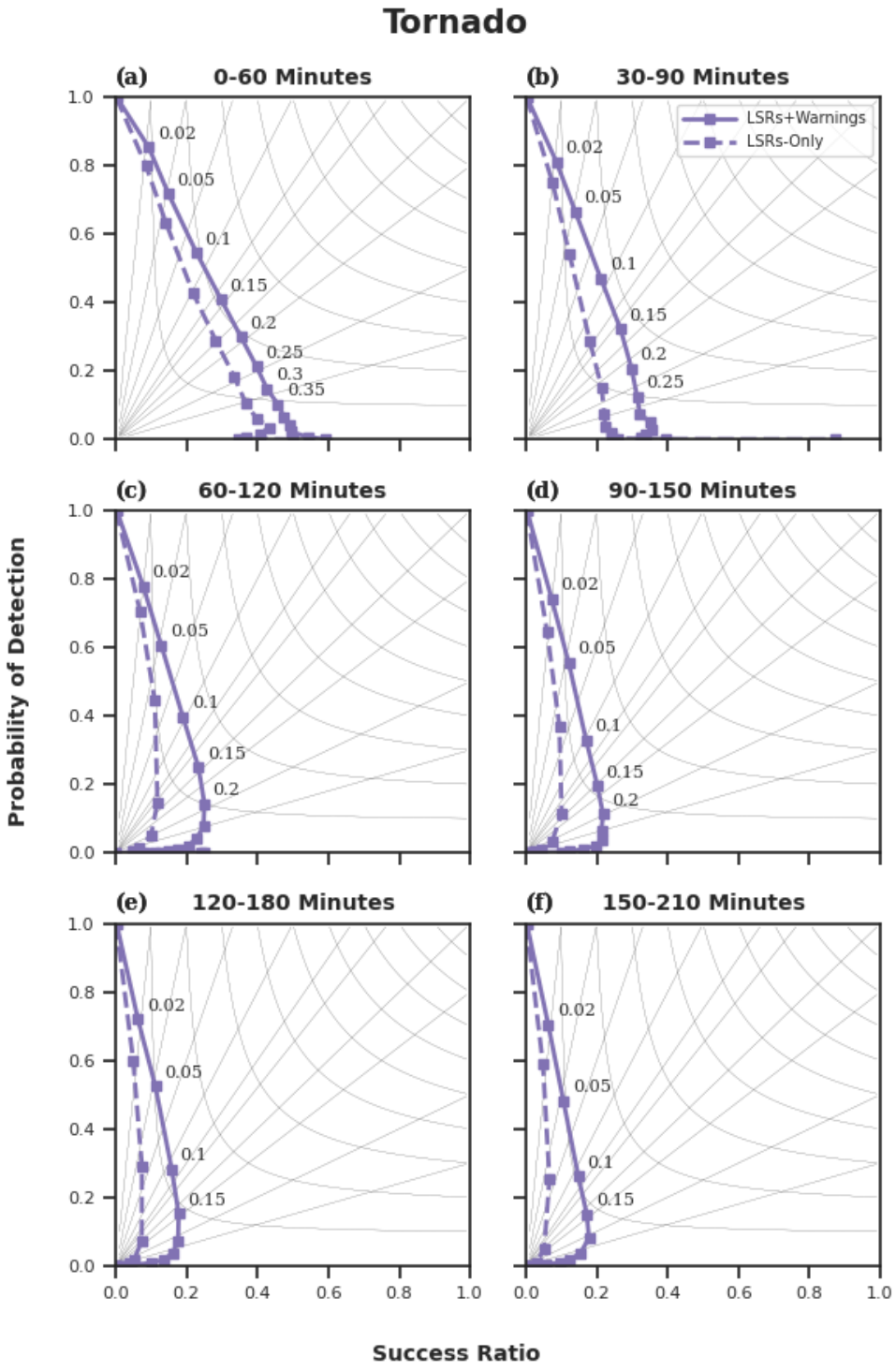

Figure 3.14: As in fig. 3.10 but for tornadoes.

3.3 Impact of TORP without ProbSevere

After observing the minimal impact that the inclusion of TORP had, an additional experiment was run to further analyze the impact of TORP as an observation-based dataset compared to ProbSevere. To do this, a model trained using WoFS and TORP (referred to as "TORP Only") to see whether TORP is less valuable to WoFS-PHI than ProbSevere or if it provides similar value as ProbSevere to WoFS-PHI. In either case, it would be expected to not improve WoFS-PHI, but understanding the reason for why TORP did not add much value is important for future uses of TORP and similar products. Since this experiment was conducted after it was determined that training on LSRs and warnings leads to higher skill than training on LSRs-only, the TORP only model here is trained on LSRs and warnings.

Removing PS3 data results in a significant decrease in BSS for hail forecasts in the $0-60$ and $30-90$ minute forecast periods (Fig. 3.15a). BSS for hail is still lower without PS3 for the $60-120$ and $90-150$ minute forecast periods, but the BSS confidence interval overlaps with the confidence interval of the reference model (Fig. 3.15a). BSS without PS3 appears to be approximately equal to the reference model with PS3 in the $120-180$ and $150-210$ minute forecast periods (Fig. 3.15a).

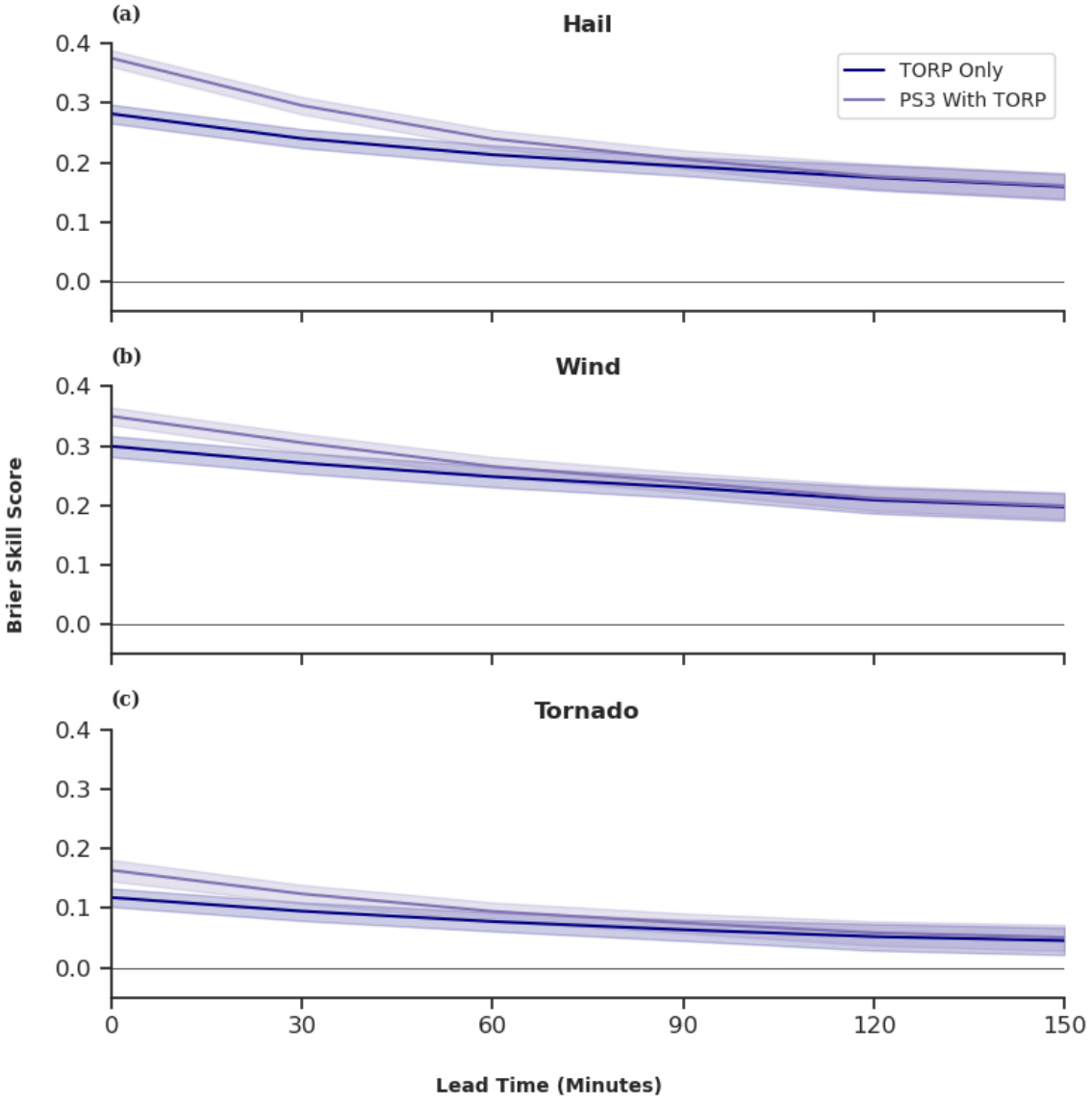

Figure 3.15: As in fig. 3.8 (hail), except both models were trained on LSRs and warnings. The dark blue line represents the model trained on TORP and no PS, and the light blue line represents the model trained on PS3 and TORP.

Hail forecast reliability, resolution, and sharpness is nearly unchanged for probabilities up to about 70% for $0-60$, $30-90$, and $60-120$ minute periods, and there is decreased resolution and sharpness for forecasts above 70% (Fig. 3.16a-c). For the $90-150$ minute period, resolution and sharpness decrease less, and mostly in

the $50-80\%$ range (Fig. 3.16d). For the $120-180$ and $150-210$ minute forecast periods, there is nearly no change in reliability, resolution, or sharpness at any probability level (Fig. 3.16e,f). The significant decrease in BSS for hail forecasts that is associated with the removal of PS3 is more evident in the performance diagrams than attributes diagrams, and it appears to be most driven by a decrease in POD, though there is some decrease in SR as well (Fig. 3.17). As with BSS, the biggest decrease is in the $0-60$ minute period, the decrease is small by the $90-150$ period, and there is nearly no difference between the model with and without PS3 in the $120-180$ and $150-210$ minute periods (Fig. 3.17).

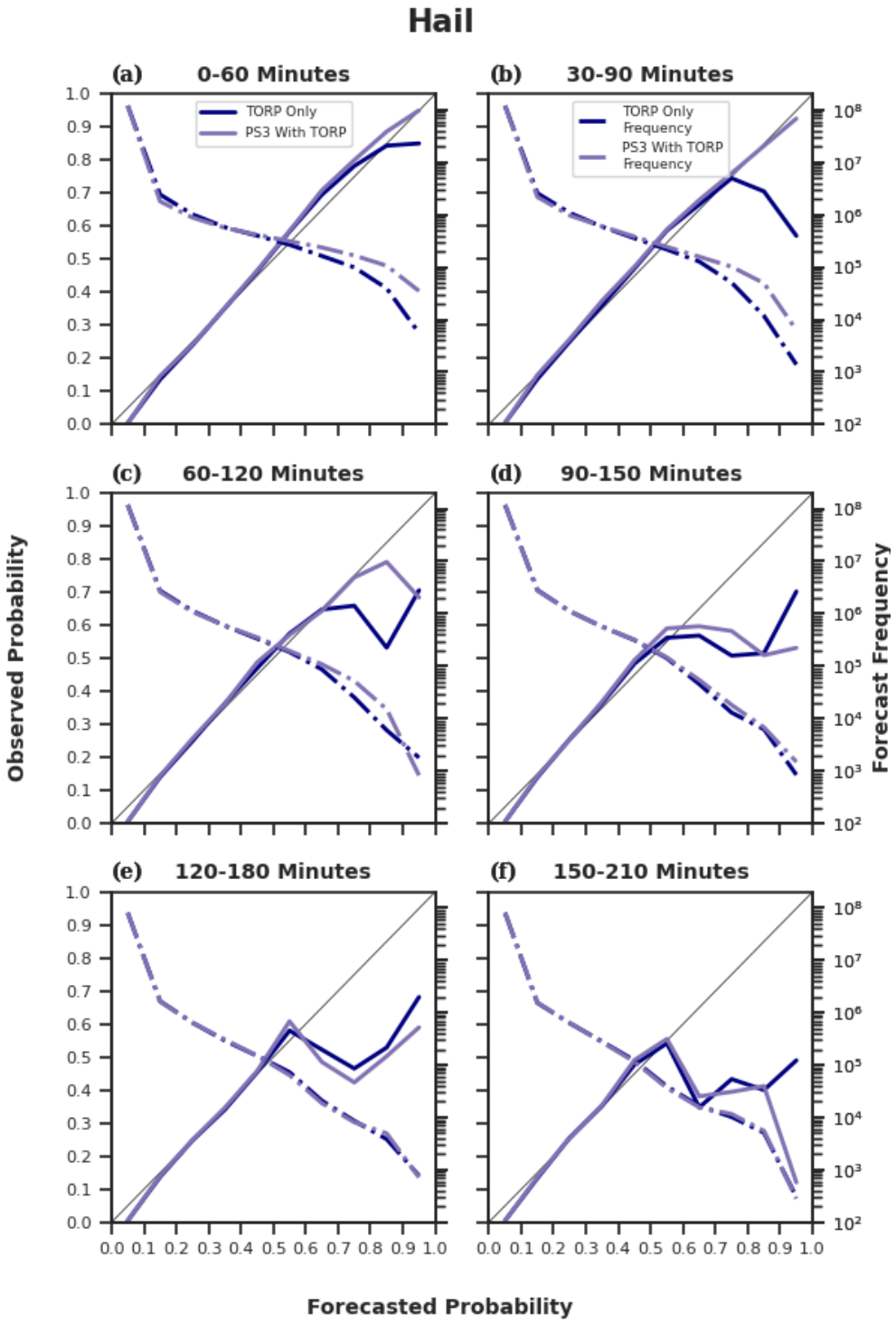

Figure 3.16: As in fig. 3.9 (hail), except both models were trained on LSRs and warnings. The dark blue line represents the model trained on TORP and no PS, and the light blue line represents the model trained on PS3 and TORP.

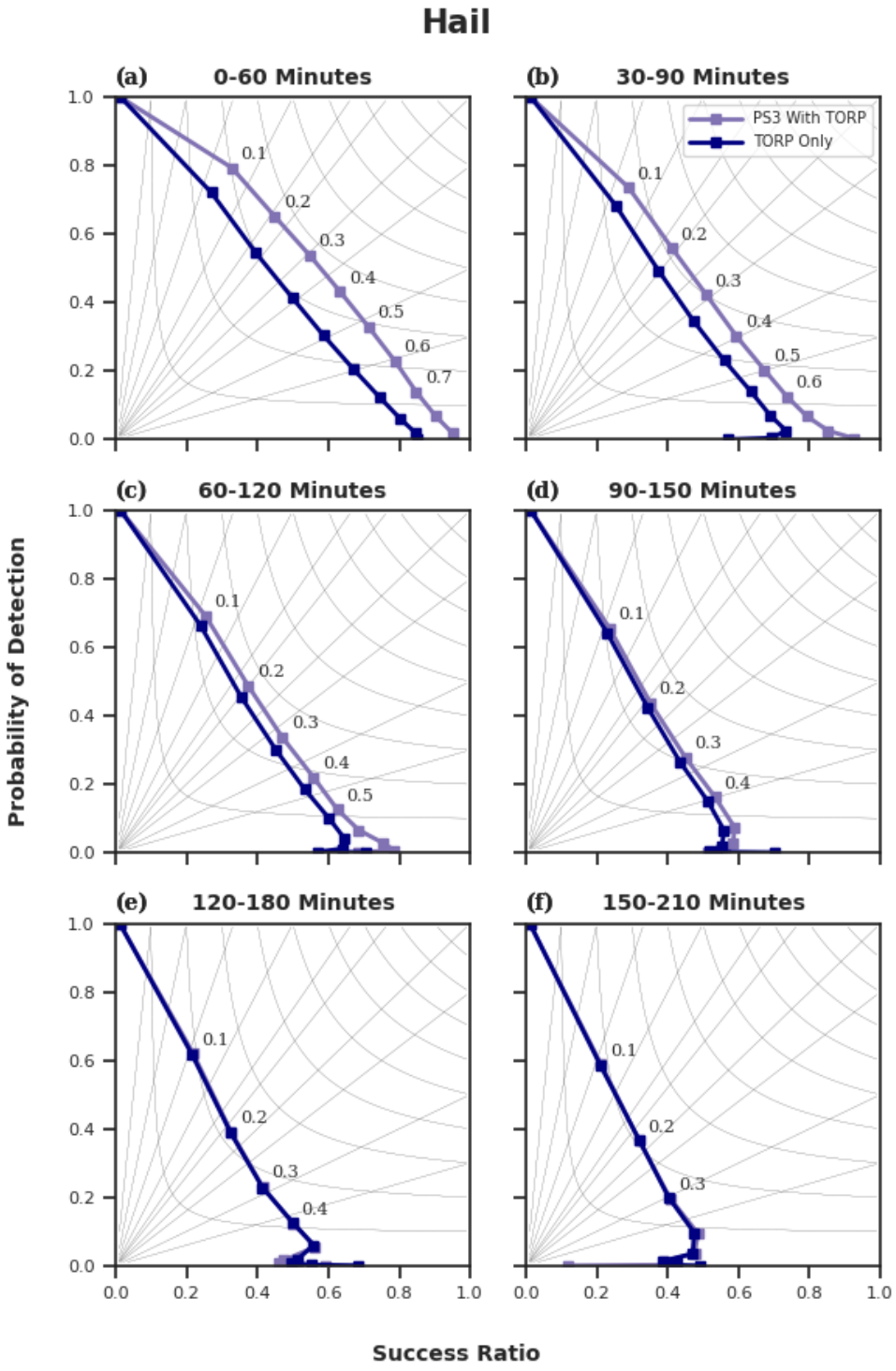

Figure 3.17: As in fig. 3.10 (hail), except both models were trained on LSRs and warnings. The dark blue line represents the model trained on TORP and no PS, and the light blue line represents the model trained on PS3 and TORP.

A similar pattern to hail is seen for wind forecasts when PS3 data is removed from the predictor dataset, but the magnitude of influence from PS3 is not as large. There is still a significant drop in BSS in the $0-60$ and $30-90$ minute periods associated with removing PS3 data, but the magnitude of the $0-60$ minute drop in BSS is about 0.05 for wind forecasts (Fig. 3.15b) compared to almost 0.1 for hail forecasts (Fig. 3.15a). For lead times of 60 minutes and longer, the drop in BSS associated with removing PS3 predictors is approximately equal between wind and hail forecasts. Interestingly, when PS3 data is removed, wind forecasts overtake hail forecasts to have higher BSS values at early lead times, and wind forecasts then have the highest BSS of all hazards at all lead times (Fig. 3.15b).

There is nearly no change to reliability or resolution at any lead time when PS3 is removed for wind forecasts (Fig. 3.18). The only potential exceptions would be a slight signal for lower resolution for forecasts greater than 80% in the $30-90$, $120-180$, and $150-210$ minute periods (Fig. 3.18b,e,f). There is also decrease in sharpness for the $0-60$ and $30-90$ minute periods, especially for forecasts greater than 80% (Fig. 3.18a,b). There is overall less of an impact on the attributes diagrams for wind forecasts by removing PS3 than there was on hail forecasts (Figs. 3.16, 3.18).

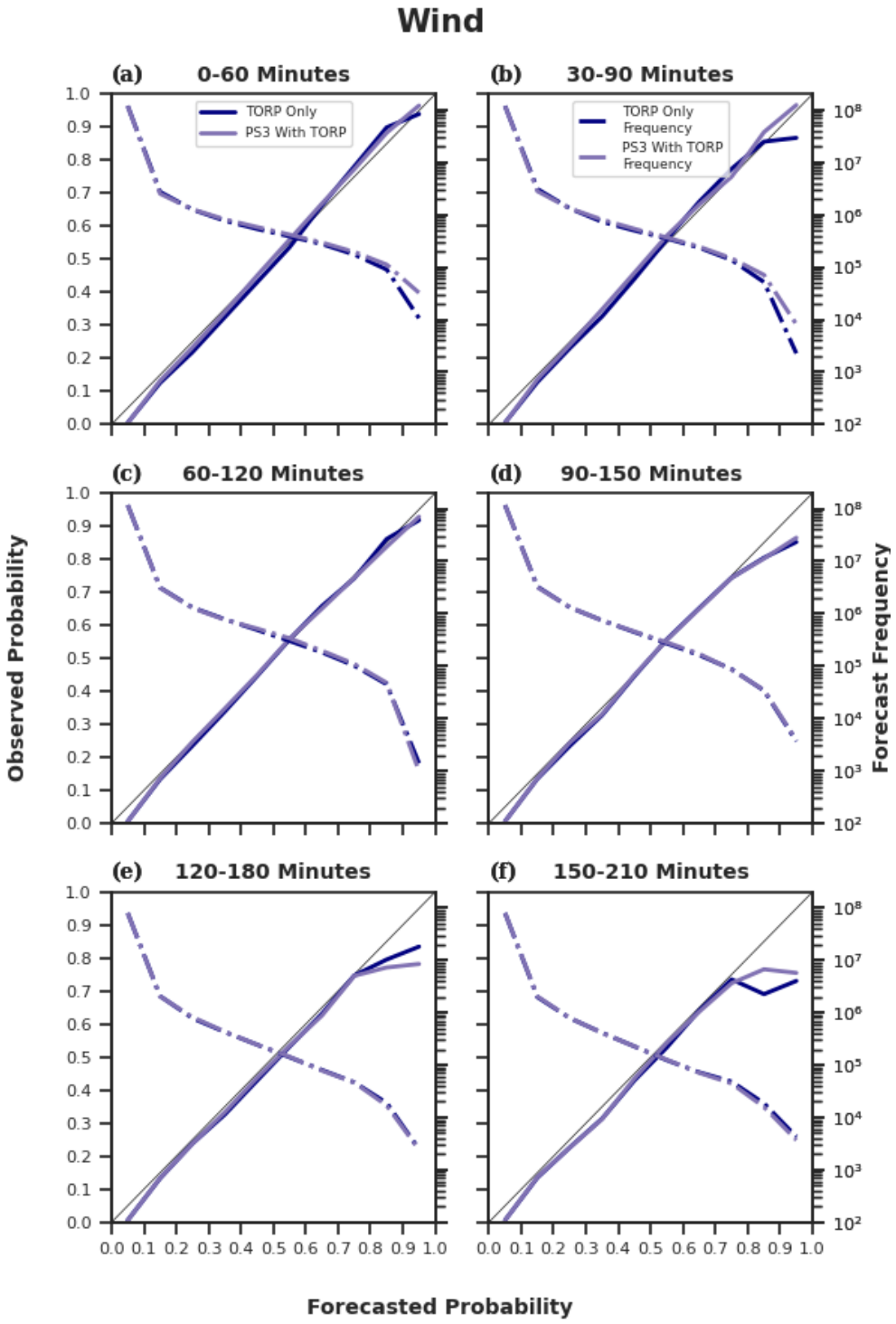

Figure 3.18: As in fig. 3.16 but for wind.

Similarly, there is less of an impact from removing PS3 on performance diagrams for wind than there was for hail. However, the pattern of most of the performance diagram change for hail forecasts being related to a drop in POD rather than SR remains the same fore wind forecasts (Fig. 3.19). The magnitude of the decrease in POD is smaller for wind forecasts, though. Consistent with the BSS pattern, the largest downward shift of the performance diagram from removing PS3 occurs in the $0-60$ minute period with lesser downward shifts occurring in the $30-90$, $60-120$, and $90-120$ minute periods, and there is no change to the performance diagrams in the $120-180$ and $150-210$ minute periods (Fig. 3.19). Removing PS3 has overall less impact on wind than hail forecast quality.

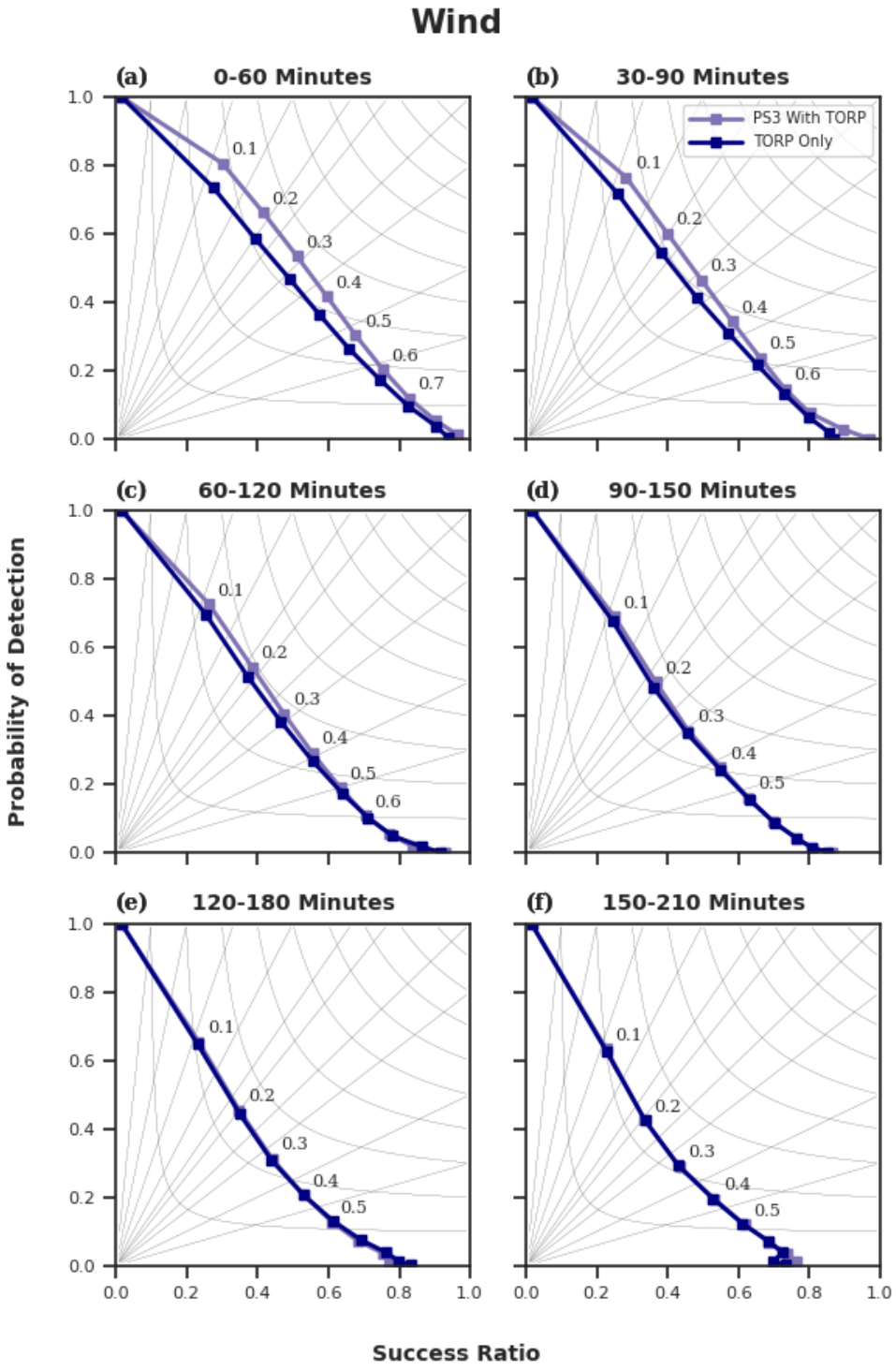

Figure 3.19: As in fig. 3.17 but for wind.

Removing PS3 from the predictor dataset has a similar impact on BSS for tornado forecasts as for wind forecasts. However, contrary to both hail and wind forecasts, the only significant drop in BSS occurs in the $0-60$ minute period, and the BSS confidence intervals for the reference and non-PS3 models overlap for lead times of 30 minutes and longer (Fig. 3.15c).

The relationship between removing PS3 and the reliablility of tornado forecasts is more complicated than for hail or wind forecasts. There appears to be decreased resolution, but increased reliability, in the $0-60$ minute period for forecasts between $20-50\%$ (Fig. 3.20a). The same range of forecasts leads to a decrease in resolution and reliability in the $30-90$ minute period (Fig. 3.20b). Interestingly, removing PS3 *increases* resolution and reliability for forecasts between $40-50\%$ in the $60-120$ and $120-180$ minute periods (Fig. 3.20c,e). However, both probability ranges have only about 10^4 forecasts issued, so it is possible that this resolution change is due more to low sample size noise than a meaningful pattern. Tornado forecast sharpness decreases in the $0-60$ minute period when PS3 is removed, but it is relatively unchanged for other forecast periods (Fig. 3.20a).

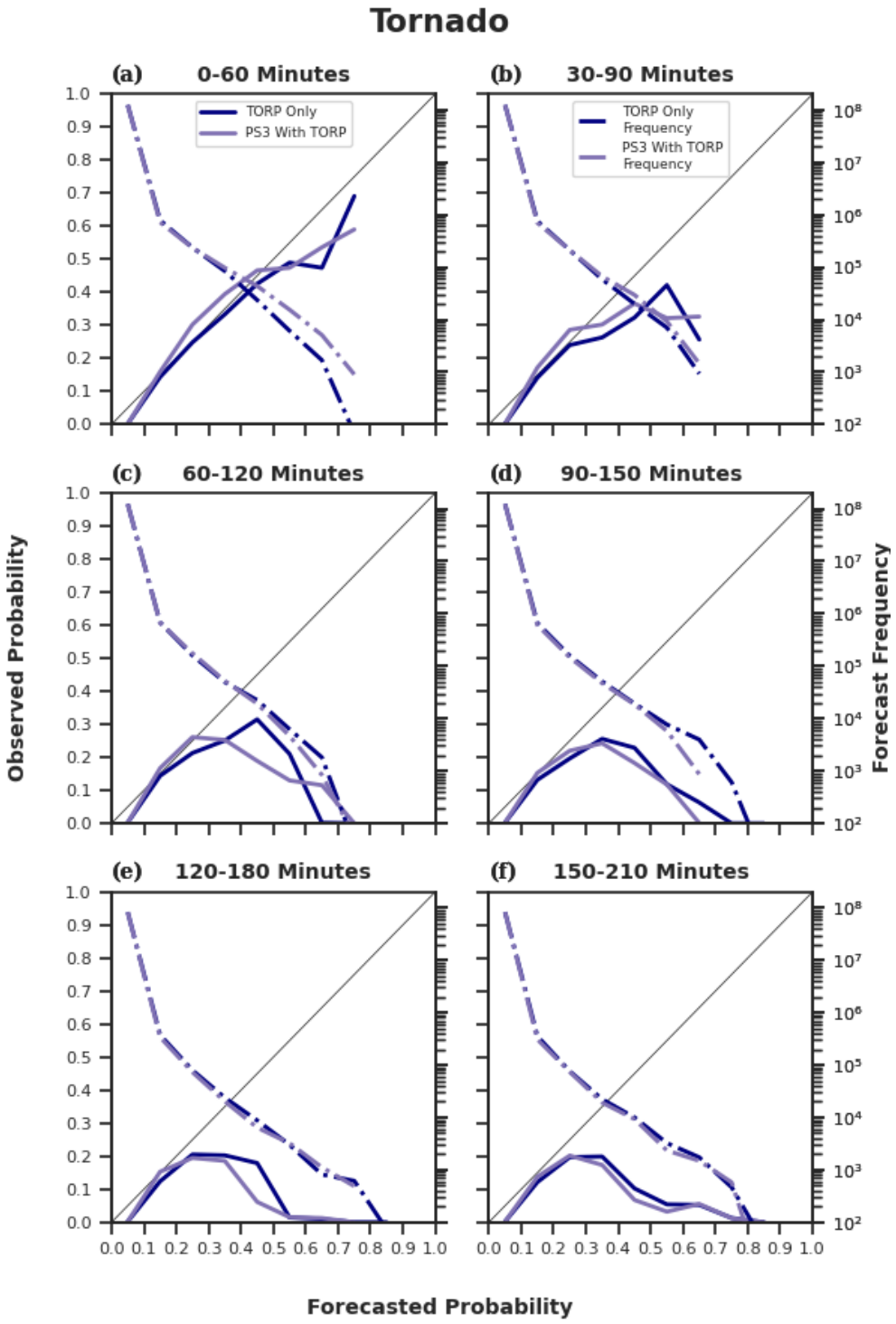

Figure 3.20: As in fig. 3.16 but for tornadoes.

The overall pattern in the performance diagrams for tornadoes is similar to wind forecasts with smaller changes in area under the curve associated with removing PS3 at early lead times disappearing (mostly) by the $90-150$ minute period (Fig. 3.21). Much of the change in area is driven by decreased PODs for lower probability thresholds, but much of the area change for higher probabilities is driven by decreased SRs, especially in the $0-60$ and $30-90$ minute periods (Fig. 3.21). The importance of the changes in SR is an effect not seen for hail and wind forecasts. It is interesting that despite tornado forecasts seeing the most improvement from the change from PS2 to PS3, the effect of removing PS3 (and not replacing it with PS2) is similar to wind forecasts and less than hail forecasts. It is possible that this is due to the fact that tornado forecast performance is worse overall, so the same magnitude of change to the lower BSS or area under the performance diagrams associated with tornado forecasts is proportionally larger than the same change to the higher BSS and area under the performance diagrams associated with hail and wind forecasts.

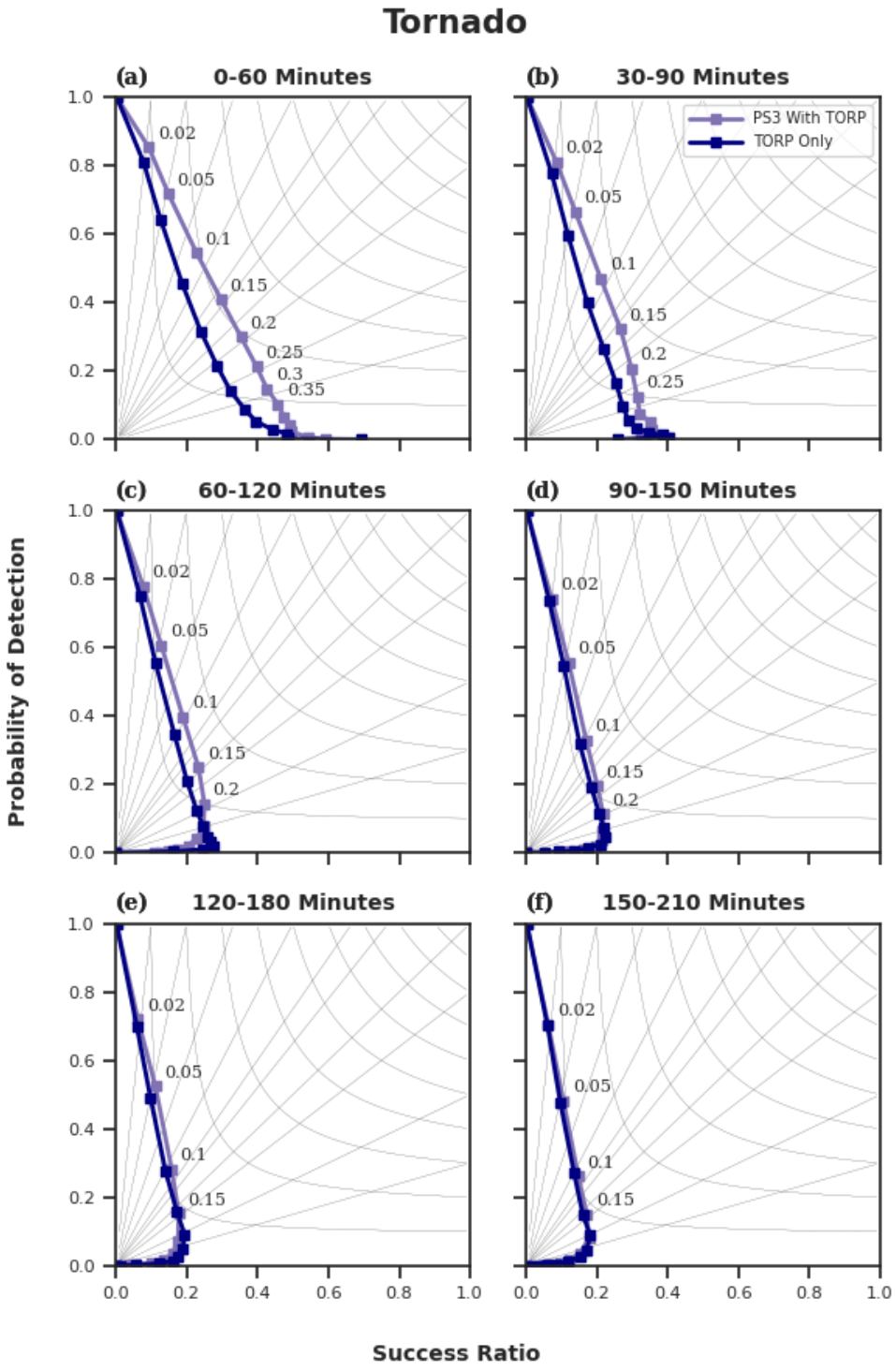

Figure 3.21: As in fig. 3.17 but for tornadoes.

Chapter 4

Explainability

4.1 Relative Tree Interpreter (RTI)

RTI assesses predictor importance within an RF model. Here, it is used to answer questions about whether different models learn different predictor relationships. As discussed in chapter 2, like predictors were grouped together, so all 269-374 (depending on the usage of ProbSevere and TORP predictors) were not analyzed individually. Twenty total predictor groups are shown, but as discussed in chapter 2, the overall contribution from WoFS, ProbSevere, and TORP are also shown. Thus, only the top 17 or 18 (depending on the inclusion of ProbSevere and TORP) individual predictor groups are shown.

Given that any severe weather is a "rare" event (very low climatology), the vast majority of forecasts are trivial "no" forecasts. While not all correct negative forecasts are trivial, a large majority are. Thus, hits, misses, and false alarms approximately represent "difficult" forecasts. Among, hits, misses, and false alarms, hits are the only scenario where the predictor has contributed correctly, so this analysis focuses primarily on each predictor's contribution to hits. Since the contribution to hits, misses, and false alarms is quite low, the secondary x-axis is used to enlarge the contribution to hits and misses for ease of viewing.

4.1.1 Impact of Changing the Predictor Dataset

A comparison between the PS2 without TORP and PS3 with TORP training datasets where both models are trained on LSRs-only is conducted first to evaluate the impact of using PS3 and TORP as predictors compared to PS2 only. An RTI analysis of the model that has access to both PS3 and TORP allows for an analysis of how the model uses each dataset (and each predictor within each dataset) individually. Thus, only one experiment is conducted rather than separate experiments to analyze the predictor importance for PS3 and TORP.

4.1.1.1 PS2 without TORP

4.1.1.1.1 Hail

In the baseline hail model that uses PS2 and no TORP, ProbSevere and WoFS contribute similarly to hits in the 0-60 minute forecast window (Fig. 4.1a). However, PS2 contributes more to correct negatives and is the most important dataset overall in the first lead time. This result suggests that PS2 is most used to determine areas to rule out for severe weather. As lead time increases, WoFS predictor contributions to both hits and overall importance increases, and PS2 predictor importance decreases (Fig. 4.1). This pattern is consistent with Loken et al. (2022b, in review), and it makes sense that the observational PS2 data would be used more for the shorter term forecasts. However, even though PS2 is most important at early lead times, the most important individual predictor field at all lead times, both overall and in terms of contribution to hits, is WoFS $2-5$ km UH, and this is also consistent with known hail indicators in NWP models (Kain et al. 2008, 2010; Sobash et al. 2011; Naylor et al. 2012; Sobash et al. 2016, 2019). WoFS $2-5$ km UH is also the most important predictor field for wind and tornado forecasts (Figs. 4.2, 4.3). In fact,

at lead times of 30 minutes and beyond, $2-5\ km$ UH contributes more than all PS2 predictors combined. Other important WoFS predictors are also generally consistent with literature in surface pressure, lightning flash extent density (Williams et al. 1999; Schultz et al. 2009; Jurkovic et al. 2015), updraft speed (Rosenfeld et al. 2008), and mid-level lapse rate (Johnson and Sugden 2014). Surface pressure could make some physical sense as severe convection would often be associated with forcing around a surface cyclone with lower surface pressure. However, it is perhaps more likely evidence of some overfitting.

Among PS2 predictors, PS2 hail probabilities contribute the most at all lead times except for the 150-210 minute forecasts at which the PS2 object age becomes most important among PS2 predictors, and PS2 hail probabilities are no longer in the top 18 predictor fields. It is notable that for 0-60 minute forecasts, PS2 wind probability is the next most important predictor field. However, PS2 wind probability importance decreases with lead time faster than PS2 hail probability. Additionally, latitude and longitude rank relatively highly in importance at all lead times, which is another concern for overfitting in a meteorological sense (whether to true hail climatology or to LSR reporting biases) if not a strict machine learning sense. This is especially notable given that there are no spatial convolutions of latitude and longitude, so while a predictor like PS2 hail probabilities is the result of aggregating 10 predictors, latitude and longitude is only the aggregation of 2 predictors.

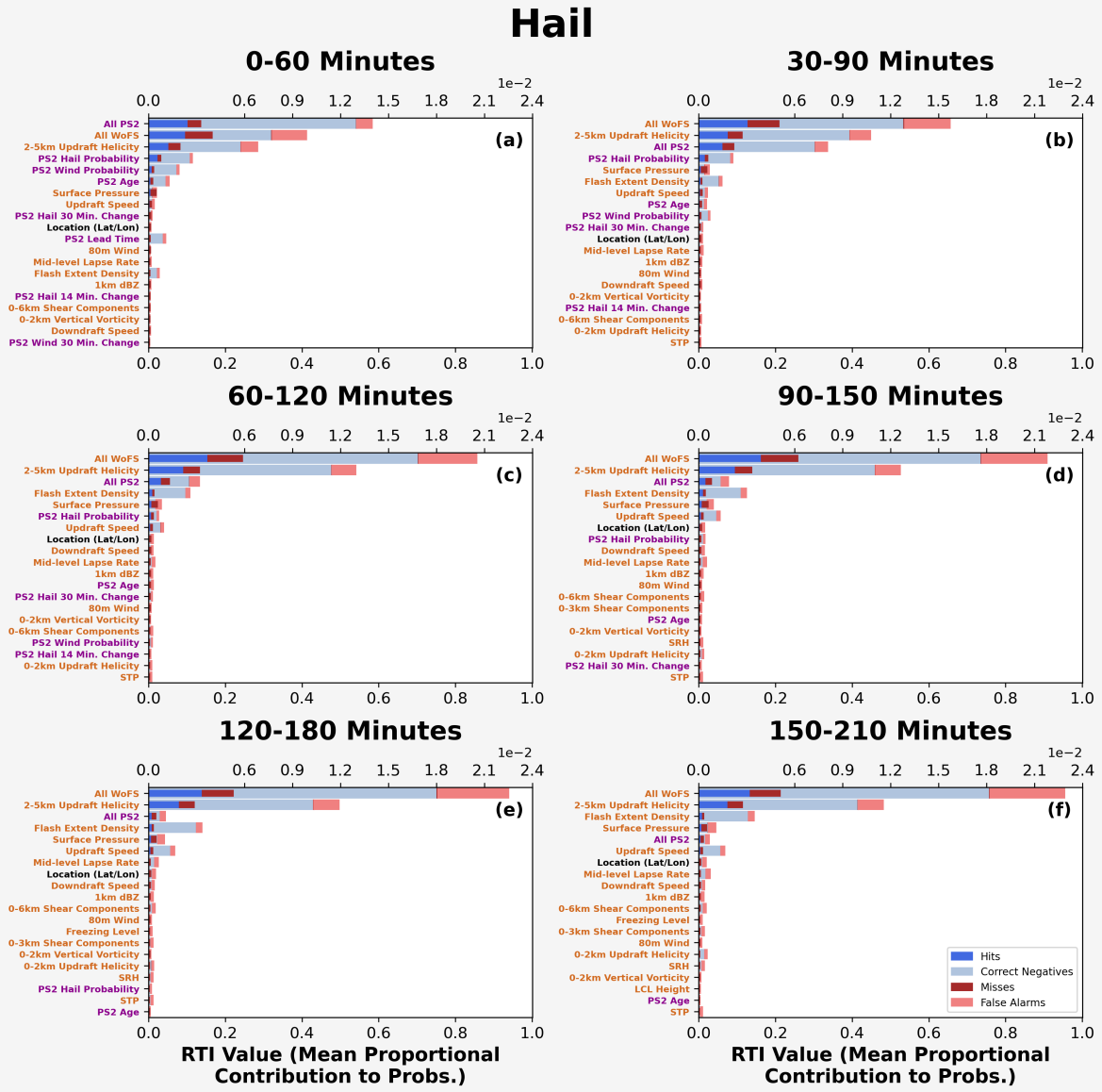

Figure 4.1: Relative tree interpreter (RTI) values for severe hail forecasts are shown for each lead time using the PS2 with no TORP model. Brown labels correspond to WoFS predictors, purple labels correspond to PS predictors, and black labels are used for latitude and longitude (geographical predictors). The order of predictors is determined by their contribution to hits (blue bars), and each predictor's contribution to hits is plotted with the contribution to misses (red) with respect to the top x-axis to increase the visibility of otherwise small contributions. The full contribution to hits, misses, false alarms (light red), and correct negatives (light blue) are underlayed on the contribution to hits and misses alone, and the total contribution uses the bottom x-axis.

4.1.1.1.2 Wind

For wind, WoFS predictors are more important than PS2 predictors at all lead times (Fig. 4.2). Although, the pattern of increasing WoFS predictor importance and decreasing PS2 predictor importance with lead time remains as the 0-60 minute forecast has nearly evenly split contributions overall and to hits from PS2 and WoFS. Important WoFS predictors include: $2-5$ km UH , 1 km dBZ (decreasing importance with lead time), surface pressure, 80 m and 10 m wind speed, updraft speed, lightning flash extent density (increasing importance with lead time), $0-2$ km vorticity and UH, and downdraft speed (increasing importance with lead time). Lightning (Williams et al. 1999; Schultz et al. 2009), UH (Kain et al. 2008, 2010; Sobash et al. 2011; Naylor et al. 2012; Sobash et al. 2016, 2019), and updraft speed (Kain et al. 2010) are consistent with hail, but lower tropospheric wind speeds are more unique to wind forecasting (Kucher and D. 2006). Among PS2 predictors, PS2 wind probability is most important only at the first two lead times before PS age is most important, and PS2 wind probabilities are not in the top 18 predictors at all at 120 and 150 minutes of lead time (Fig. 4.2e,f). Latitude and longitude are similarly important at all lead times as with hail. Wind LSRs are known to have a bias towards a higher frequency in the southeastern US (Smith et al. 2013; Edwards et al. 2018), and it is one of the weaknesses of using LSRs as a target dataset. The importance of latitude and longitude for wind LSR forecasts could be reflective of this bias.

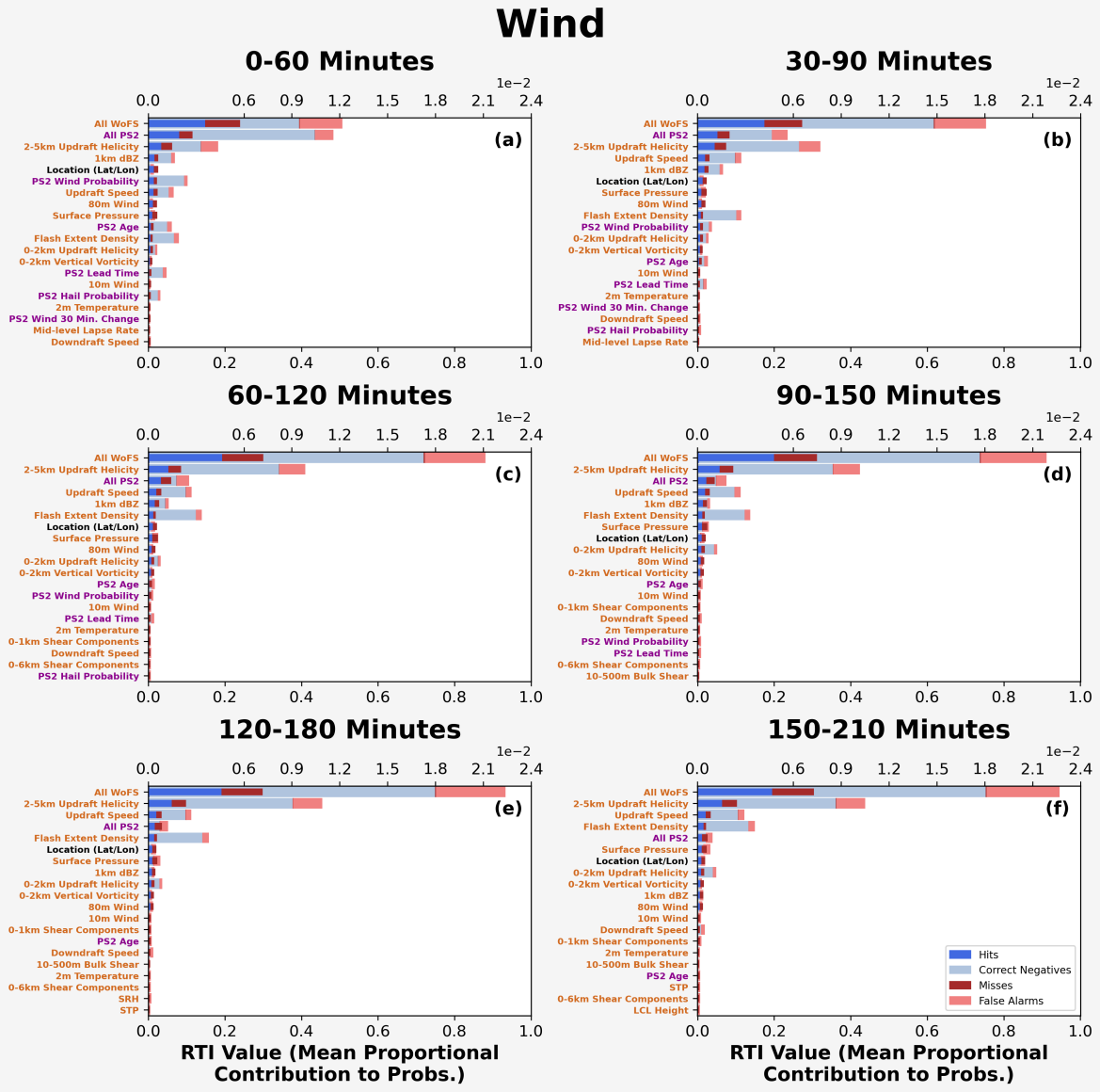

Figure 4.2: As in fig. 4.1, but for wind.

4.1.1.1.3 Tornadoes

For tornadoes, a similar pattern to wind of WoFS predictors contributing more to hits than PS2 predictors is present (Fig. 4.3). However, PS2 predictors have overall more importance in the 0-60 minute period and approximately equal importance in the 30-90 minute period (Fig. 4.3a,b). Important WoFS predictors include: $2-5\ km$

and $0-2$ km UH (increasing with lead time; Kain et al. 2008, 2010; Sobash et al. 2011; Naylor et al. 2012; Sobash et al. 2016, 2019), STP (Thompson et al. 2003), surface pressure, $0-2$ km vorticity (Sobash et al. 2019), updraft speed (increasing with lead time; Kain et al. 2010), SCP (increasing with lead time; Thompson et al. 2003). These predictors are also similar to what was found in Loken et al. (in review) and are consistent with the literature (Kain et al. 2008, 2010; Sobash et al. 2011; Naylor et al. 2012; Sobash et al. 2016, 2019). Similar to wind, PS2 tornado probabilities are the most important PS2 predictor at 0-60 minutes and 30-90 minutes, but the object age is most important afterwards. However, PS2 tornado probabilities are among the top 18 predictors even in the 150-210 minute forecast period. Latitude and longitude are less important for tornado predictions compared to hail and wind, especially at early lead times, but geographical dependence approaches a similar level to the other hazards at later lead times.

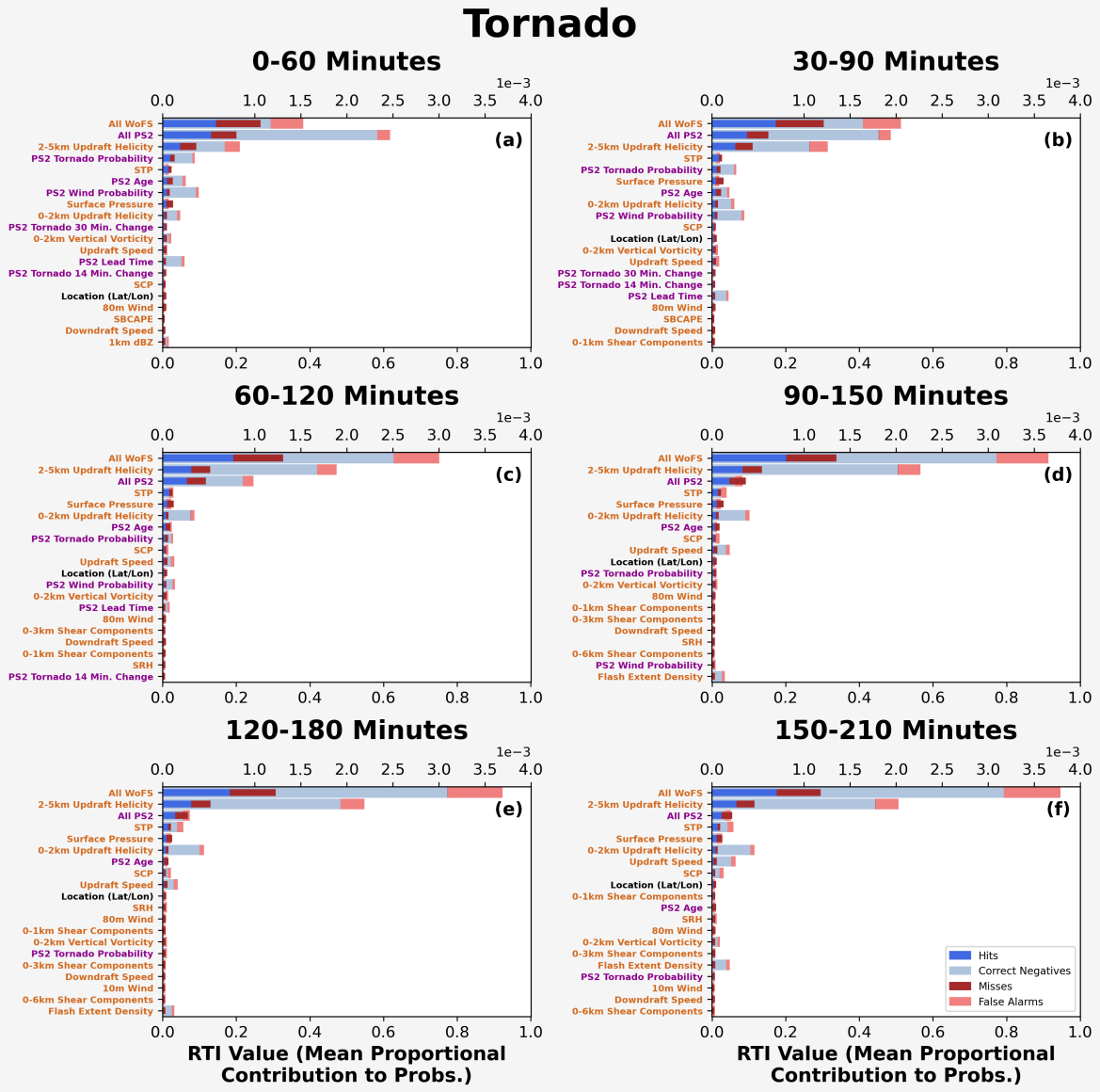

Figure 4.3: As in fig. 4.1, but for tornadoes.

4.1.1.2 PS3 with TORP

4.1.1.2.1 Hail

When PS3 and TORP are used, many of the same WoFS predictors are important for hail (Fig. 4.4), and the same pattern appears where WoFS (PS3) predictors increase (decrease) with lead time. Similarly, $2-5$ km UH is the most important

predictor at all lead times. This is unsurprising given that there were no changes made to the WoFS predictors. PS3 wind probabilities are less important than PS2 wind probabilities were in the baseline model. However, PS3 tornado probabilities are more important than they were with PS2. PS3 hail probabilities remain the top predictor through the first five lead times before dropping out of the top 17 predictors for the 150-210 minute forecasts. After 150 minutes, PS3 object age becomes the most important predictor group from ProbSevere, just like in the baseline model that used PS2. TORP does not contribute much to the forecasts, but the one predictor that is used is the maximum reflectivity within the storm associated with the TORP object (hereafter referred to as simply "maximum reflectivity"). This predictor becomes less important with lead time but does appear in the top 17 predictors for the first 3 lead times. This makes sense as its importance should decrease similarly to PS3 as they are both observational data sources, and reflectivity is also known to be correlated with severe hail (Auer Jr. 1994). Latitude and longitude are also slightly less important than for hail forecasts from the baseline model, especially at early lead times.

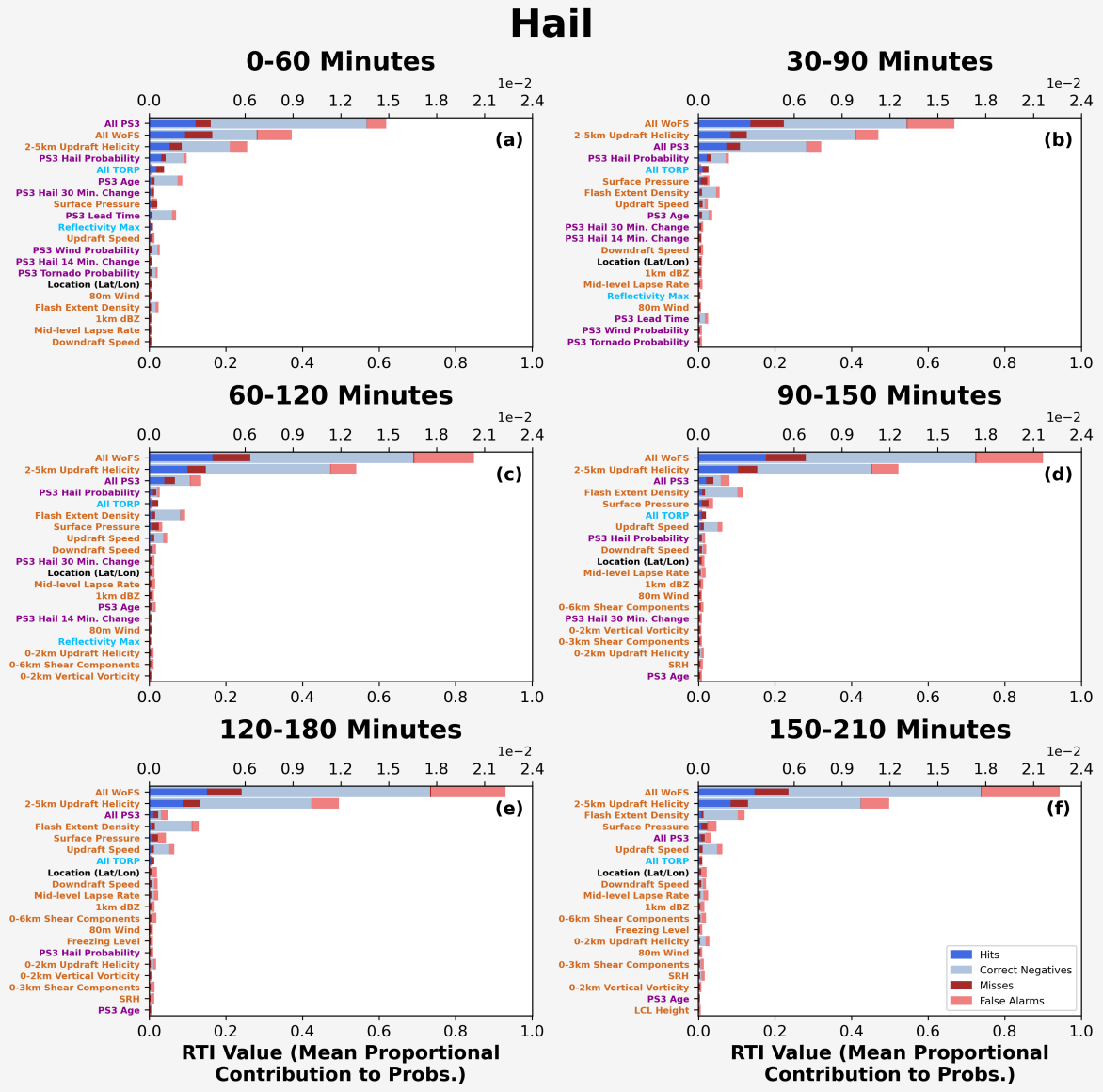

Figure 4.4: As in fig. 4.1, but using the PS3 with TORP model. Cyan labels represent predictors from TORP.

4.1.1.2.2 Wind

For wind, there is an increase in the importance of PS3 predictors (compared to PS2 importance in the baseline model) in the first four lead times, though it is most dramatic at the earliest lead times (Fig. 4.5). Additionally, there is an increase in usage of PS3 wind data for predicting correct negatives in the 0-60 minute forecast

period. Conversely, there is a large drop in usage of WoFS predictors for correct negative prediction. Similar to hail, many of the same WoFS predictors are important. However, the importance of latitude and longitude is relatively unchanged when using PS3 and TORP compared to PS2 only, unlike with hail forecasts.

PS3 wind probabilities have similar importance at early lead times to what PS2 wind probabilities had in the baseline model. However, at later lead times, PS3 wind probabilities are more important than PS2 wind probabilities were in the baseline model. Additionally, PS3 wind probabilities are the most important PS3 predictor field at all lead times. This is in contrast to PS2 wind probabilities that were the most important PS2 predictor only in the 0-60 and 30-90 minute windows. TORP predictors are also not very important. Further, the extrapolation lead time of the TORP objects is the most important predictor field in the 0-60 minute period, which does not have much physical basis. This lends credence to the assertion that TORP data is not particularly important to WoFS-PHI. After the 0-60 minute period, TORP does not appear in the top 17 predictor fields.

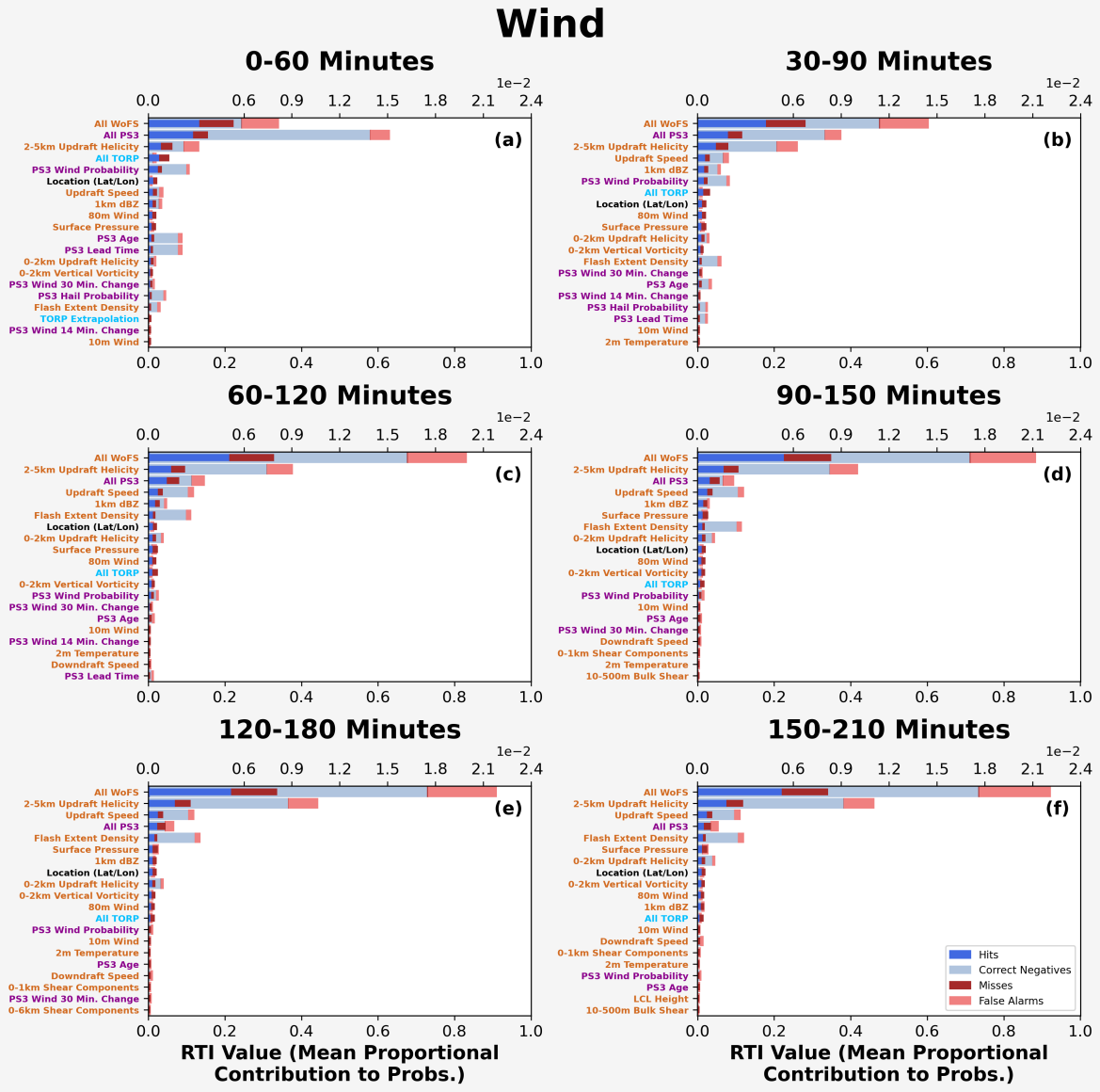

Figure 4.5: As in fig. 4.4, but for wind.

4.1.1.2.3 Tornadoes

There is minimal overall difference in the importance of PS3 data relative to PS2 data for tornadoes (Fig. 4.3). While PS3 becomes more important than WoFS for hit contribution (it was already greater for overall contribution), the actual change in PS3/WoFS predictor importance is small. However, PS2 and WoFS were very similar

in hit contribution, so a slight shift towards PS data when PS3 was introduced was enough to flip the order. Similar to the other hazards, the pattern of important WoFS predictors is mostly unchanged. Among PS3 predictors, PS3 wind probabilities are less important than PS2 wind probabilities had been. It is also encouraging to note that the most important TORP predictor was TORP probability, which makes sense for tornado prediction. However, after the 0-60 minute forecast period, no TORP predictors appear in the top 17. Although, similar to hail and wind forecasts, the aggregate effect of all TORP predictors remains present in the top 17 at all lead times.

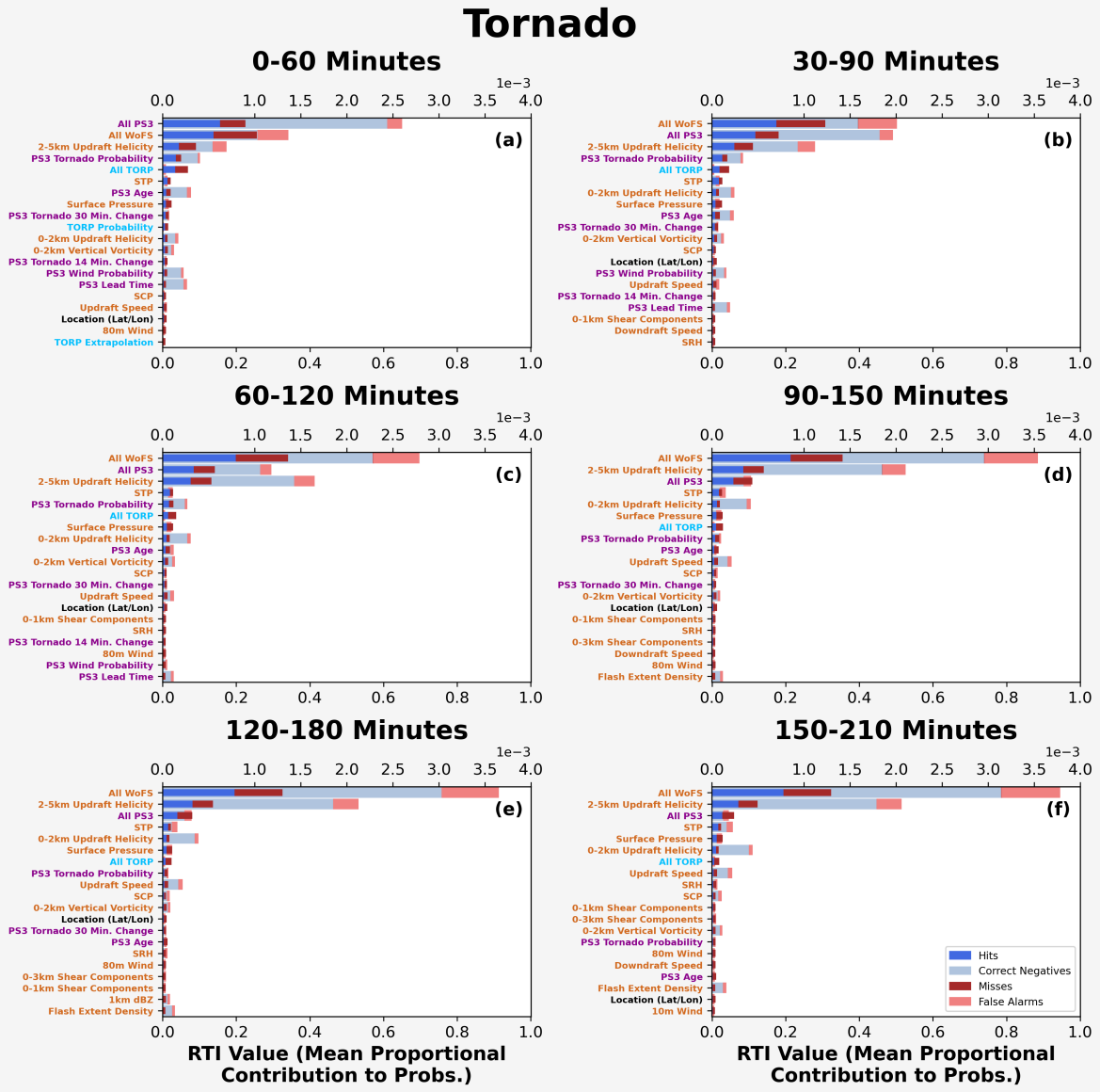

Figure 4.6: As in fig. 4.4, but for tornadoes.

4.1.2 Impact of Including Warnings in the Target Dataset

In this section, the RTI analyses of the model trained using PS3 and TORP predictors with a target dataset of LSRs and warnings is compared against the RTI analyses of the model trained with PS3 and TORP predictors with a target dataset of LSRs-only from the previous section (Figs. 4.4, 4.5, 4.6). This comparison will show

whether WoFS-PHI learns different predictor relationships when forecasting LSRs and warnings instead of just LSRs.

4.1.2.1 Hail

The biggest impact for hail forecasts is that there is an increased contribution to hits from both WoFS and PS3 (Fig. 4.7). This makes sense since the climatology of events will necessarily be higher than when only LSRs are used as events. Thus, there are more opportunities for a hit to occur, so the fraction of forecasts that contribute to a hit will increase. Similarly, the contribution to misses increases as well. However, it is notable that the contribution to misses does not increase nearly as much as the contribution to hits.

This suggests that the RF model is more skillful when forecasting LSRs and warnings than when forecasting LSRs alone since despite the increased climatology of events, the fractional contribution to misses does not increase much despite the increased opportunity for missed events that results from a higher climatology. This is consistent from the verification results that showed significant skill improvements when training on warnings in addition to LSRs. Some of this result could be due to the increased sharpness of the LSR and warning trained model (and thus likely more instances of increased probabilities than decreased probabilities). However, a model was trained to make 2-hour long forecasts (not shown) that had increased sharpness (especially at early lead times), but it had proportional increases in contributions to hits and misses in the RTI analysis, unlike with the LSR and warning trained model compared to the LSR-only model.

Relative predictor importance does not appear to change much for hail when trained on LSRs and warnings compared to LSRs-only (Figs. 4.4, 4.7). The most important WoFS predictors are generally the same. Both tornado and wind PS3

probabilities are slightly more important with warnings included as events. There is also some decrease in the importance of latitude and longitude compared to the LSR-only trained model. The importance of TORP predictors slightly decreases as well with the warnings-trained model.

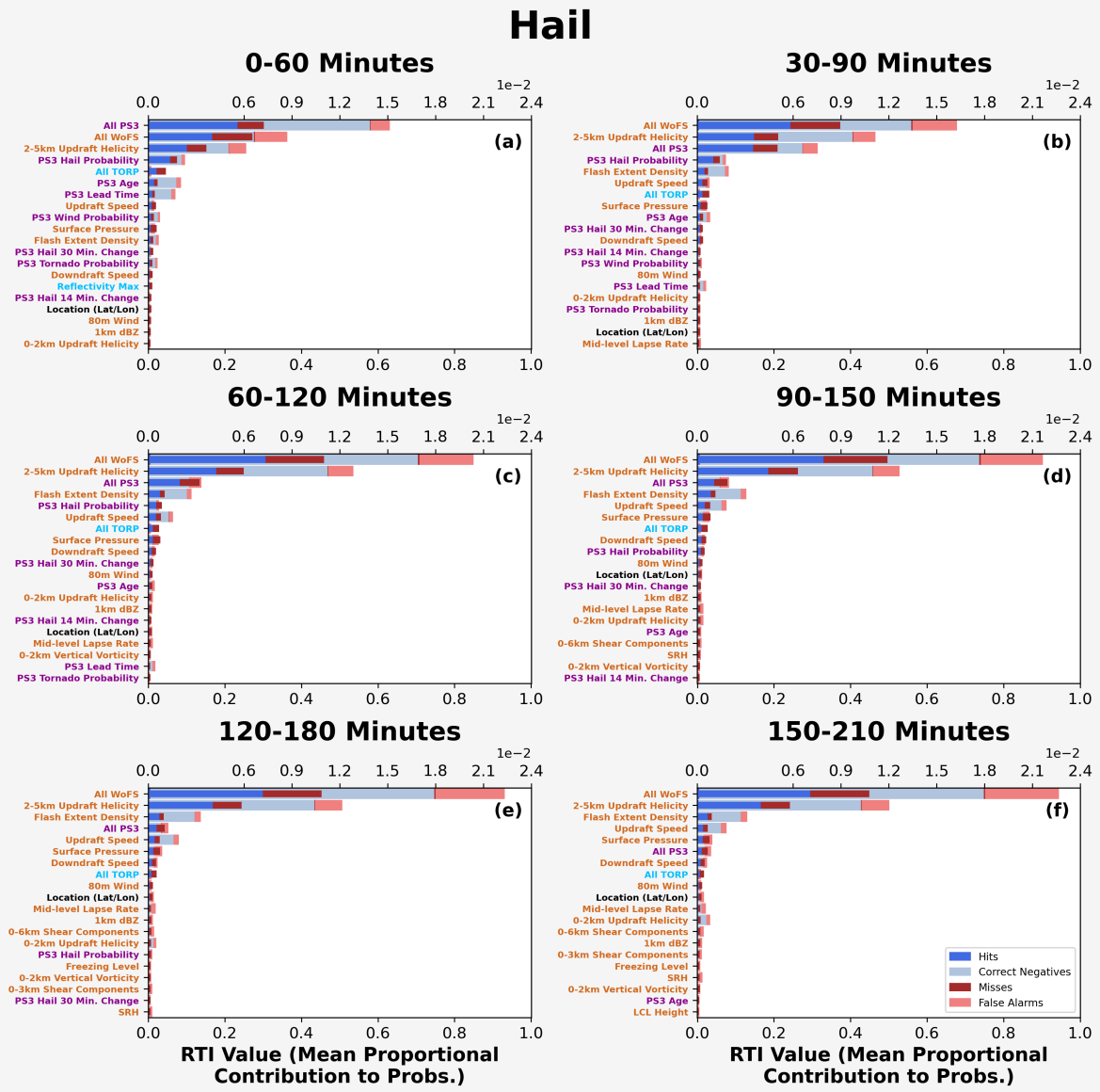

Figure 4.7: As in fig. 4.4, but for the model trained on LSRs and warnings.

4.1.2.2 Wind

Similar to hail forecasts, there is an increased contribution to hits from all predictors at all lead times when warnings are included in the target dataset (Fig. 4.8). Perhaps the most obvious effect that is unique to wind forecasts from including warnings in training is the decreased importance of latitude and longitude (Figs. 4.5, 4.8). This trend is especially noticeable at early lead times. Among WoFS predictors, there is a shift towards less 1 km dBZ being less important and flash extent density being more important, especially at early lead times. There is minimal other change in the most important WoFS predictors.

Training on LSRs and warnings led to increases in the importance of hail and tornado PS3 probabilities for wind forecasts (Figs. 4.5, 4.8). This impact is also greatest at the early lead times, which makes sense as PS3 is most important at the early lead times. It also makes sense for PS3 hail probabilities to be more important for wind forecasts given that many severe thunderstorm warnings have tags for both wind and hail, so forecasts for wind warnings should be coupled with hail more than forecasts for wind LSRs.

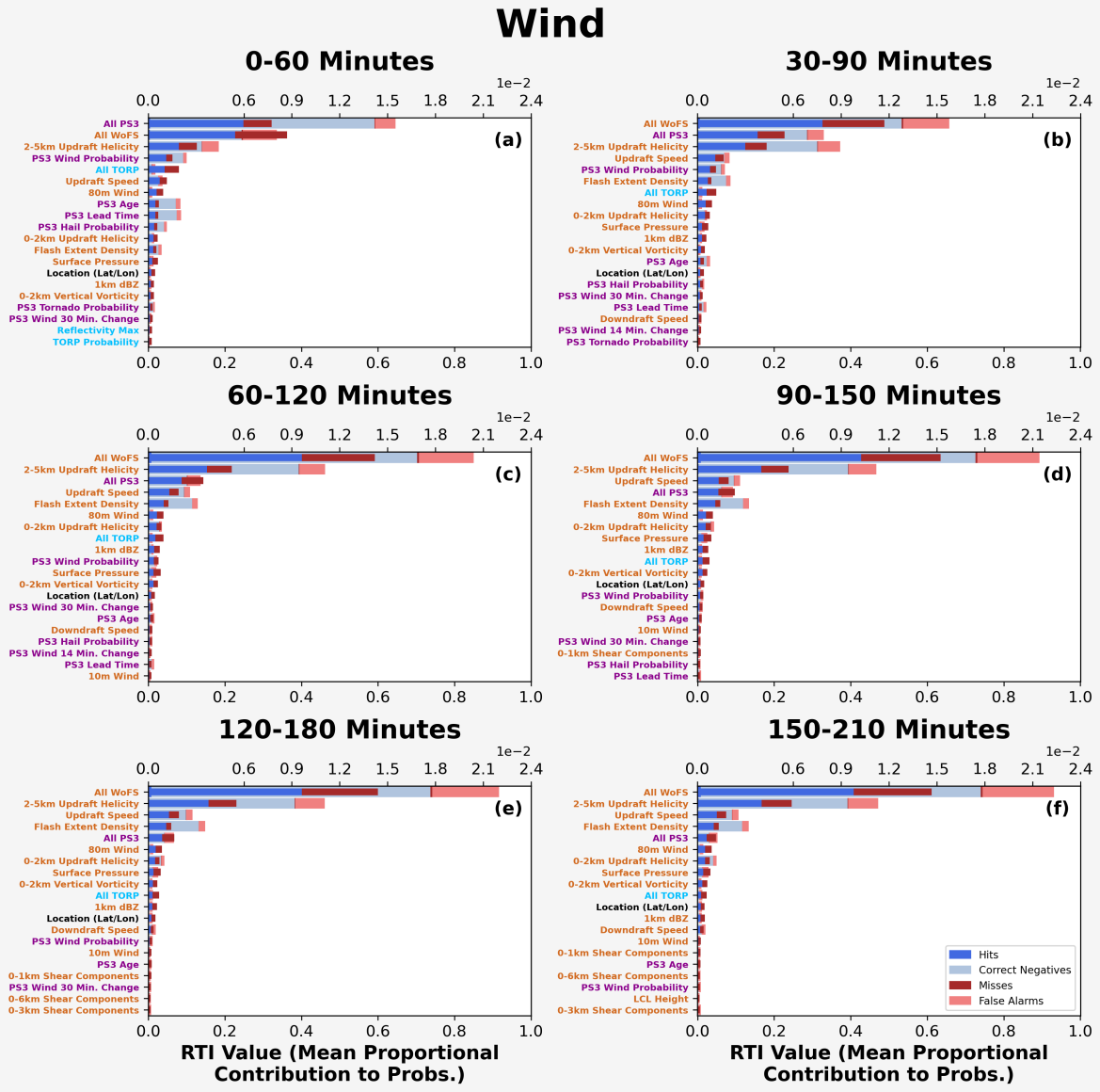

Figure 4.8: As in fig. 4.7, but for wind.

4.1.2.3 Tornadoes

Tornadoes similarly see increased contributions to hits from all predictors when warnings are included in training (Fig. 4.9). There is overall minimal change in WoFS predictor importance when training on warnings for tornado forecasts (Fig. 4.9). There is a slight decrease in the importance of $0-1$ km and $0-3$ km shear

when warnings are included, and perhaps a slight increase in the importance of $0-2$ km UH and vorticity. There is also an increased importance for 1 km dBZ (at all lead times) and flash extent density (especially at later lead times) with the model trained on warnings. Among PS3 predictors, there is a slight increase in the importance of PS3 wind probabilities at early lead times, and a decrease in the use of PS3 tornado probabilities at later lead times. It is encouraging (although perhaps somewhat surprising) that treating "tornado possible" tagged severe thunderstorm warnings as tornado warnings did not lead to higher PS3 hail and wind probability importance for tornado forecasts. Interestingly, there is also an increase in the importance of latitude and longitude for the 150-210 minute forecasts, but it is mostly unchanged for the other forecast windows.

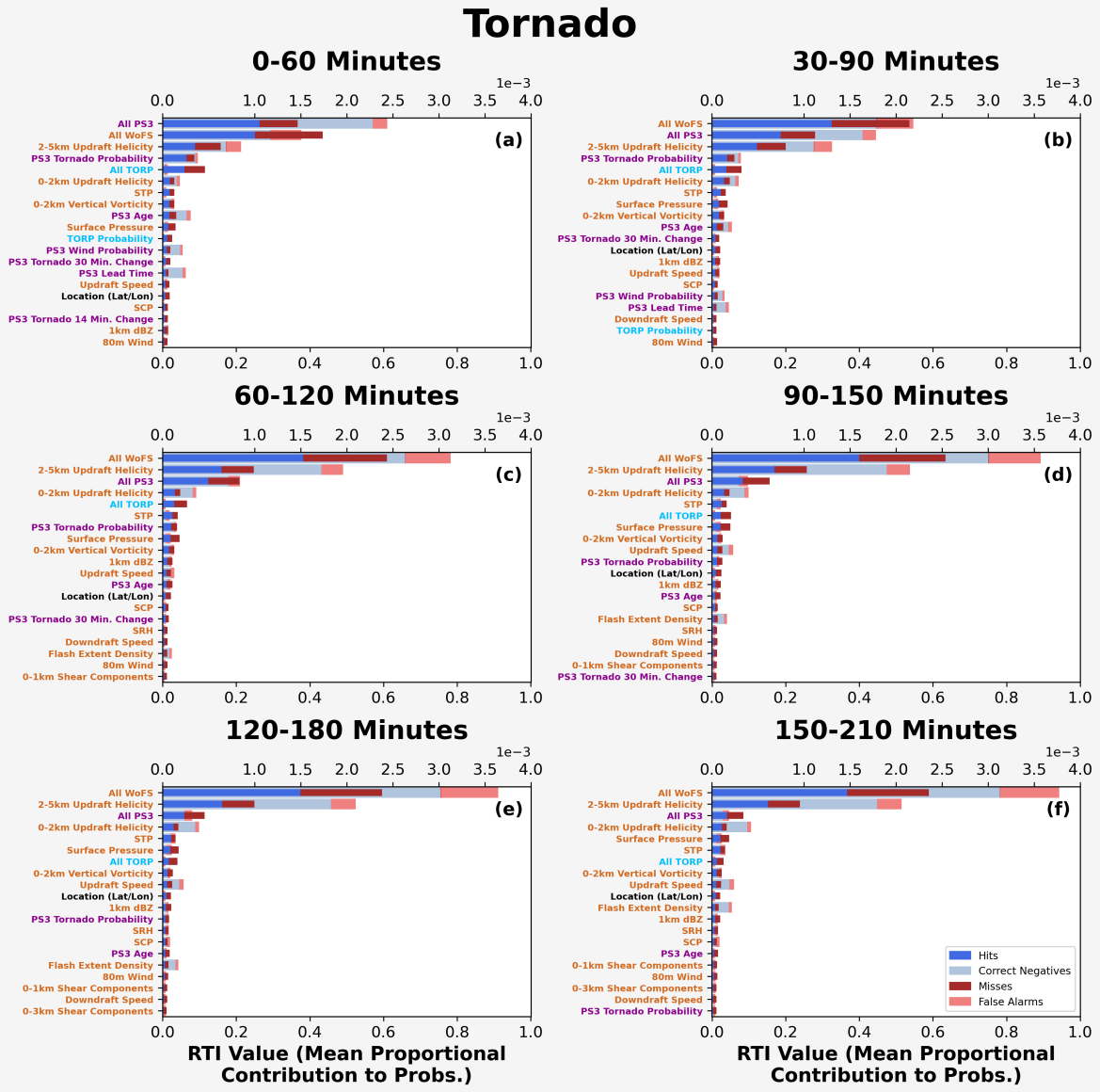

Figure 4.9: As in fig. 4.7, but for tornadoes.

4.1.3 Impact of TORP without PS3

While it was expected that the model that used WoFS and TORP without PS3 would have worse verification, this model was created to analyze which TORP predictors are important to WoFS-PHI without being washed out by PS3 predictors.

4.1.3.1 Hail

When PS3 predictors are removed from training, it is clear that the vast majority of the void left by PS3 data is filled with contributions from WoFS predictors for hail (Fig. 4.10). Further, it appears that WoFS $2-5$ km UH is used in place of PS3 data at early lead times. These experiments are the only situation in which WoFS $2-5$ km UH *decreases* in importance with lead time. Since overall WoFS importance is relatively unchanged with lead time, this is reflective of WoFS storm attribute fields being most important at early lead times and WoFS environment fields becoming more important with increasing lead time (although, $2-5$ km UH remains most important at all lead times).

That WoFS is more important than TORP to the forecasts suggests that these forecasts are worse due to under forecasting at early lead times with neither TORP nor $2-5$ km UH providing the spatiotemporal precision of PS3. This behavior is evident in the attributes diagrams at early lead times (Fig. 3.16a-c). However, the lack of sharpness only appears for the higher-end probabilities ($\geq 70\%$), and these situations likely arise in areas of convection that are new or poorly resolved by WoFS.

While it is not necessarily surprising that TORP does not fully fill the observational data void left by PS3, the extremely minimal usage of TORP data even when it is the only source of observational data available is somewhat surprising. However, the most important TORP predictor for hail is still maximum reflectivity (Fig. 4.10), which is encouraging that WoFS-PHI is using what would be expected to be the most relevant TORP predictor.

Also notable is that the TORP maximum reflectivity is the 2nd most important predictor field for contributing to hits whereas the 1 km dBZ from WoFS is just the 12th most important predictor for contributing to hits (Fig. 4.10). This suggests that the observed reflectivity values associated with TORP objects are more important to

WoFS-PHI forecasts than the modeled values produced by WoFS. Further, this difference is exaggerated from when PS3 data is included (TORP max reflectivity was 12th, and WoFS 1 km dBZ was 16th; Fig. 4.7). Radar reflectivity appears to be one of the tools that WoFS-PHI uses to replace the loss of PS3 hail probabilities, and it gravitates towards observed reflectivity values associated with TORP objects as opposed to modeled reflectivity associated with WoFS. There is also a small increase in the usage of latitude and longitude for hail forecasts when PS3 data is removed.

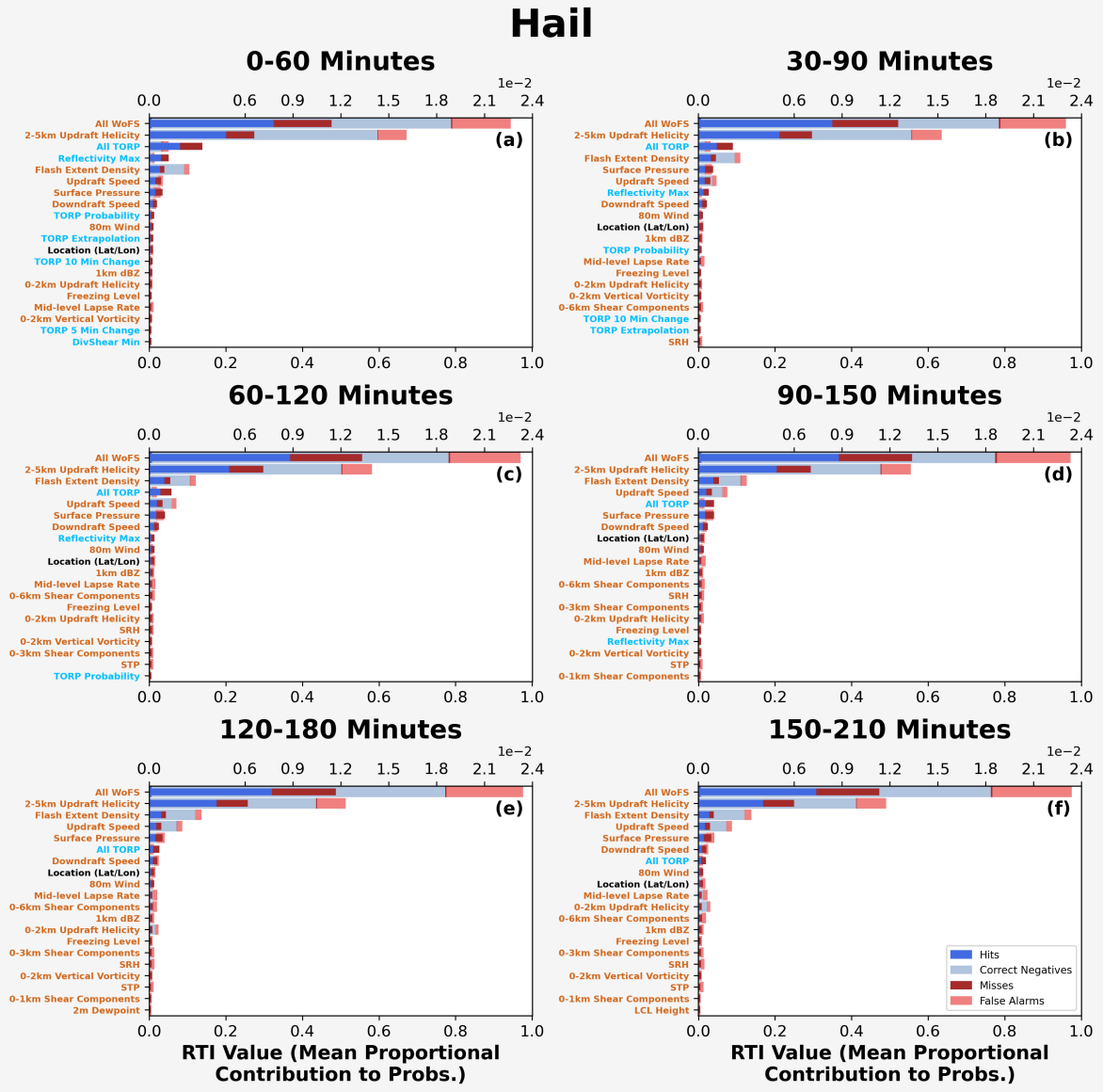

Figure 4.10: As in fig. 4.4, but for the model trained without PS data.

4.1.3.2 Wind

Similar to hail forecasts, TORP is mostly unimportant to WoFS-PHI wind forecasts even in the absence of PS3 (Fig. 4.11). Unique to wind (Fig. 4.11), there is less physical reasoning behind the TORP predictors that are most important in the RTI analysis. Maximum reflectivity is still the most important TORP predictor (5th

overall), followed by TORP probability (8th overall). Perhaps what should be the most relevant TORP predictor, maximum velocity, is just the 4th most important TORP predictor (14th overall). In fact, only TORP maximum reflectivity and TORP probability are more important than latitude and longitude.

Interestingly, while TORP maximum reflectivity was used much more than WoFS 1 km dBZ for hail forecasts (Fig. 4.10), for wind, these two predictors have nearly equal importance. This is further evidence of the fact that while maximum reflectivity is the most important TORP predictor, it is still not particularly important if it is used to a similar degree to modeled reflectivity.

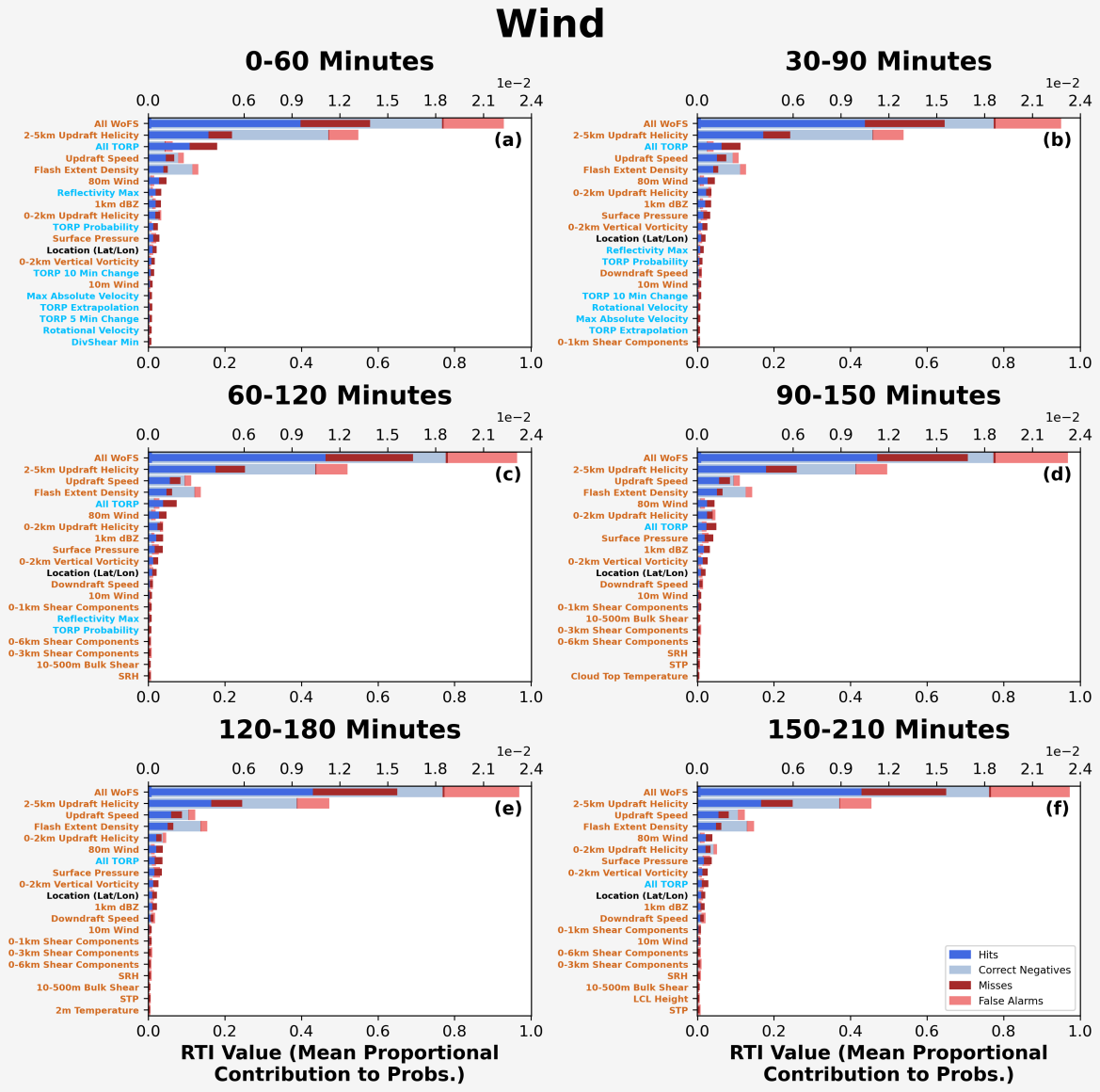

Figure 4.11: As in fig. 4.10, but for wind.

4.1.3.3 Tornadoes

TORP is also unimportant to tornado forecasts, as with hail and wind forecasts (Fig. 4.12). Interestingly, though the overall importance of TORP is similar to with hail and wind forecasts, its contribution to hits is higher than with either hail or wind forecasts. The TORP predictor fields used also make sense with TORP probability

and 10-minute probability change being the most important of the TORP predictors. Interestingly, while TORP maximum reflectivity is the 3rd most important TORP predictor field (10th overall), it is less used than WoFS 1 km dBZ (7th overall), a deviation from the trend from wind and especially hail where TORP's observational reflectivity was preferred to WoFS predicted reflectivity. The early lead times also tend to show an increased importance of latitude and longitude when PS3 data is excluded, similar to with hail forecasts.

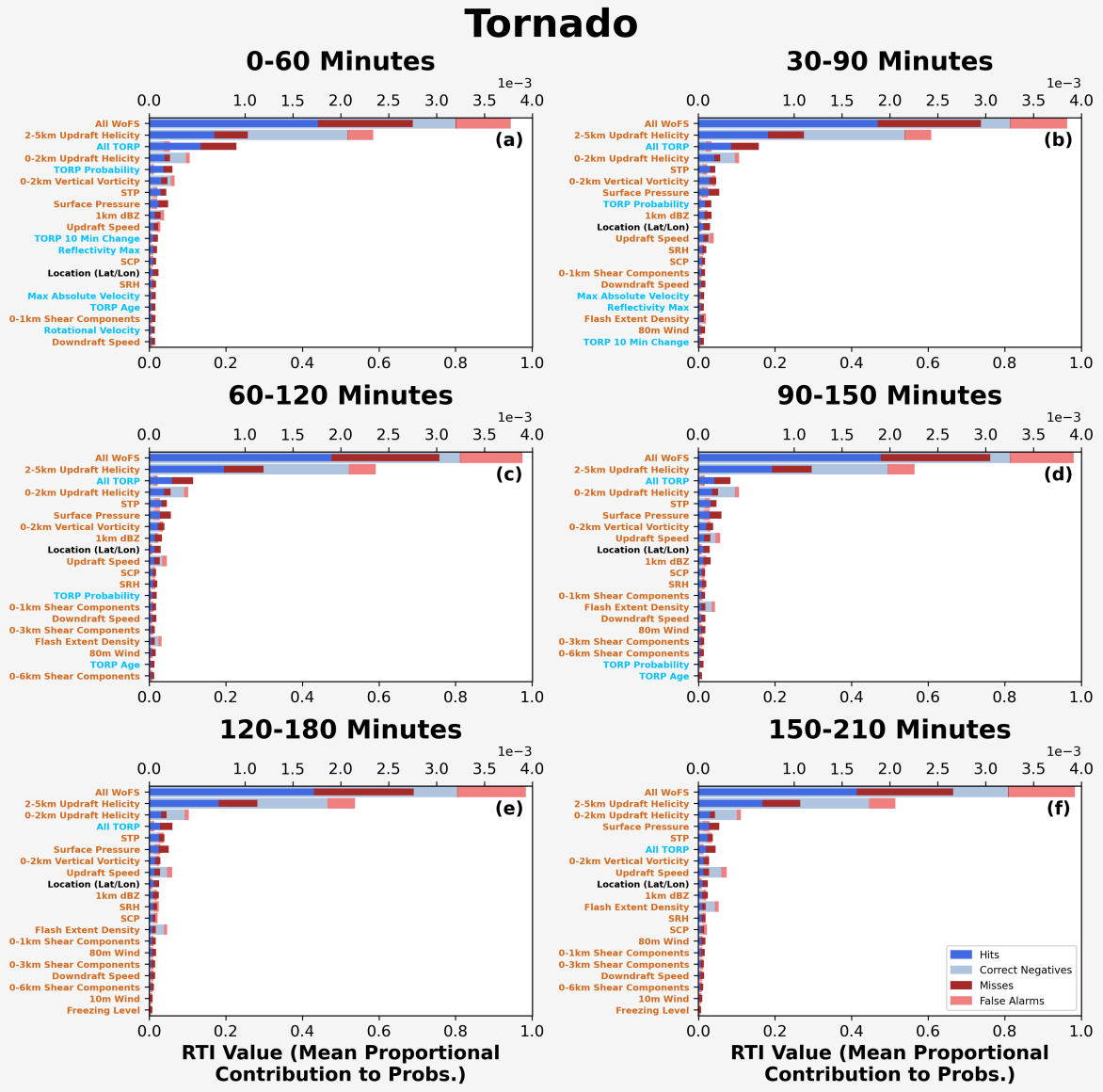

Figure 4.12: As in fig. 4.10, but for tornadoes.

It should be noted that across all hazards, WoFS filled by far the most of the void left by the removal of PS3 data. However, the noted increases in the contribution to hits from TORP data indicates that the relative increase in WoFS contribution was much greater for correct negatives rather than for hits or misses. So, while TORP is still used less than WoFS predictors (particularly $2-5$ km UH tracks) for identifying storms to raise probabilities for, this gap is smaller for difficult correct forecasts than

for the numerous trivial correct negatives that dominate the overall contribution of predictors.

4.2 Accumulated Local Effects (ALE)

Differences in model behavior from including warnings in the target dataset are clearly shown from the RTI analysis. However, changes to the predictor dataset lead to more muted effects. Thus, ALE analysis is done only for the experiments with changes to the predictor datasets. ALE analysis is useful for analyzing predictors in more depth because it allows for the influence of individual predictors at individual values to be seen. As mentioned in chapter 2, the predictors cannot be grouped together during ALE analysis as they can with RTI. Thus, for a predictor field like $2-5$ km UH that uses the 18 ensemble values, the ensemble mean, and the 15, 30, 45, and 60 km convolutions of each value (for 95 total individual predictors), the ALE value for just the 30 km convolution of the ensemble mean of $2-5$ km UH (the representative $2-5$ km UH predictor shown in the ALE analysis here) will not appear to have the same outsized effect that was seen with RTI. This lack of predictor grouping combined with the fact that hundreds of predictors are used in WoFS-PHI also means that the actual magnitudes of ALE will appear quite low. Thus, with 269-374 predictors, even an ALE value of 0.5% represents a very outsized impact from a predictor. Additionally, while ALE represents the deviation from the average prediction, the average prediction is very close to 0 given the rarity of severe weather events. Thus, sharpness differences between models can show up in ALE with higher peak ALE values across many/all important predictors indicative of higher final RF probabilities.

The goal of the ALE analysis is thus not to compare the relative effects between predictors but rather the relative effects of different predictors between models. Some comparisons between like predictors can still be done in the instances that there are equal numbers of "like" predictors. Examples would be comparisons between WoFS environmental variables or between ProbSevere probabilities for different hazards.

4.2.1 Impact of Changing the Predictor Dataset

4.2.1.1 Hail

Most important predictors seem to have increased or unchanged influence when PS3 and TORP are used compared to PS2 only. Among WoFS predictors, updraft speed and $2-5$ km UH have increased influences for higher predictor values in the $0-60$ minute period (Fig. 4.13). These changes are not extremely large, though they are large enough to be notable especially given that there was no increase in sharpness, and no changes were made to WoFS predictors. Also interesting is the pattern with longitude. While the ALE magnitude is small, there is a clear increase in average predictions west of $-96°$, and a decrease to the east. It is possible that this is due to overfitting, but given the very low ALE values, it seems that longitude is not impacting the final RF probability much. That would indicate that the effect is rather a reflection of hail LSR climatology.

Among ProbSevere predictors, there is a decrease in the influence of PS3 wind probabilities compared to PS2 wind probabilities, but there is an increase in the influence of PS3 tornado probabilities compared to PS2 tornado probabilities. However, the biggest increase in ProbSevere usage came with PS3 hail probabilities compared to PS2 hail probabilities in the $0-60$ minute period (Fig. 4.13). Given that PS3 has been shown to be more skillful than PS2 (Cintineo et al. 2024), and PS2

was most important to WoFS-PHI at early lead times (Loken et al. 2022b, in review), it is unsurprising that PS3 hail probabilities accounted for the biggest change to predictor influences.

TORP also shows an expected pattern. There is minimal effect from the maximum reflectivity until it is over approximately 50 dBZ, at which point the influence sharply increases (Fig. 4.13). This makes sense and is consistent with high reflectivity known to be correlated with hail cores (Auer Jr. 1994). Further, there is a modest influence from TORP probability, which is consistent with the modest influence of ProbSevere tornado probabilities (Fig. 4.13). Both of these influences suggest a stronger correlation between hail and tornado forecasts than between hail and wind forecasts.

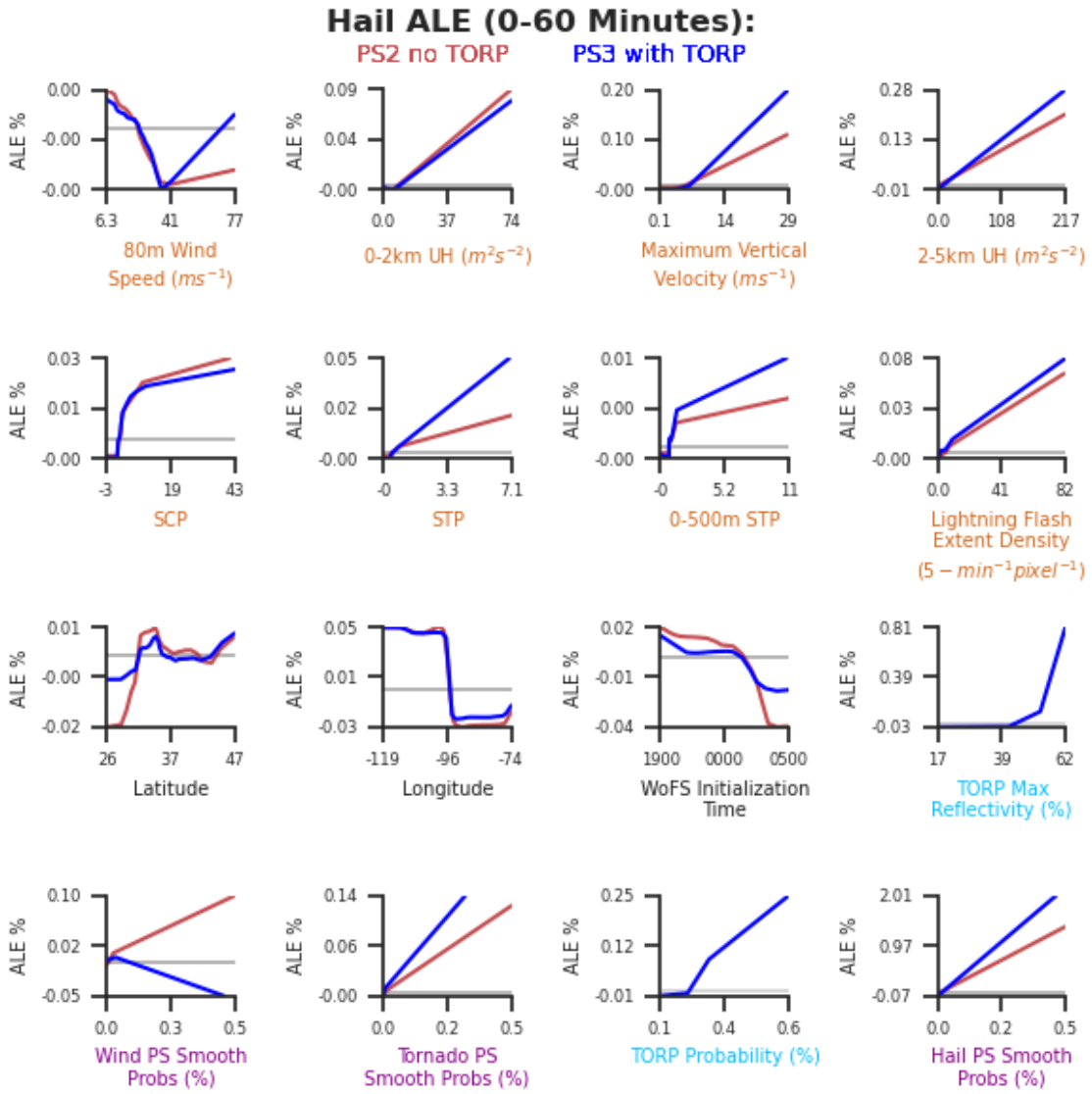

Figure 4.13: The accumulated local effect (ALE) curves of select predictors are shown for severe hail forecasts using the models trained with PS2 and no TORP (red) and PS3 with TORP (blue). The predictor's values are plotted on the x-axis, and the average deviation from the average final RF percentage is on the y-axis. Both models are trained on LSRs-only and for the 0-60 minute forecast period. Blank are placeholders for predictors that were not used in training. The same color scheme is used as with RTI plots for variable labels - brown labels/curves represent WoFS predictors, purple represents PS predictors, and black represents latitude, longitude, and the initialization time.

In the 60-120 minute period, there is no longer much of a deviation in the influence of WoFS predictors between models. However, $2-5\ km$ UH and lightning flash extent density both increase in influence at the later lead time (especially lightning) as WoFS-PHI begins to rely increasingly on WoFS predictors. Interestingly, in both models, there is an increase in the influence of longitude (the only other predictor aside from the two WoFS predictors to increase in influence compared to the 0-60 minute period), though still centered around $-96°$ (Fig. 4.14), so there could be increased overfitting as uncertainty increases. There is minimal change to ProbSevere predictor influences between PS2 and PS3 in the 60-120 minute period (Fig. 4.14), although ProbSevere is less influential for both versions compared to the 0-60 minute period, as expected. Similar patterns are also seen with TORP probability and maximum reflectivity, but the overall ALE magnitudes are much lower than in the 0-60 minute period. The generally lower predictor importances for ProbSevere and TORP are expected from both RTI results and decreased sharpness at later lead times. All predictors decrease in importance in the 120-180 minute period compared to the 60-120 minute period with continued decreasing sharpness, but the general patterns do not change much (Fig. 4.15.

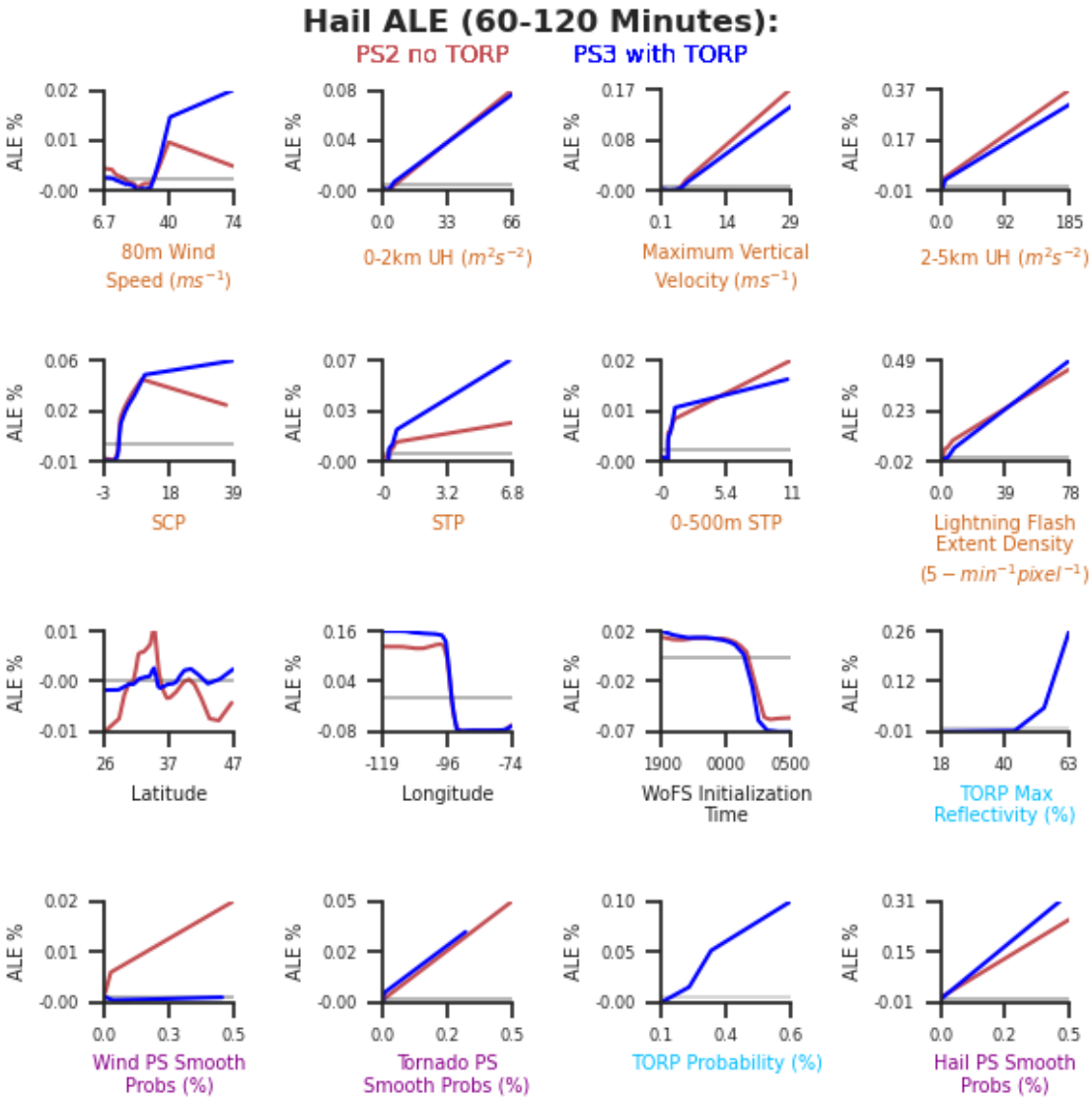

Figure 4.14: As in Fig. 4.13 but for the 60-120 minute period

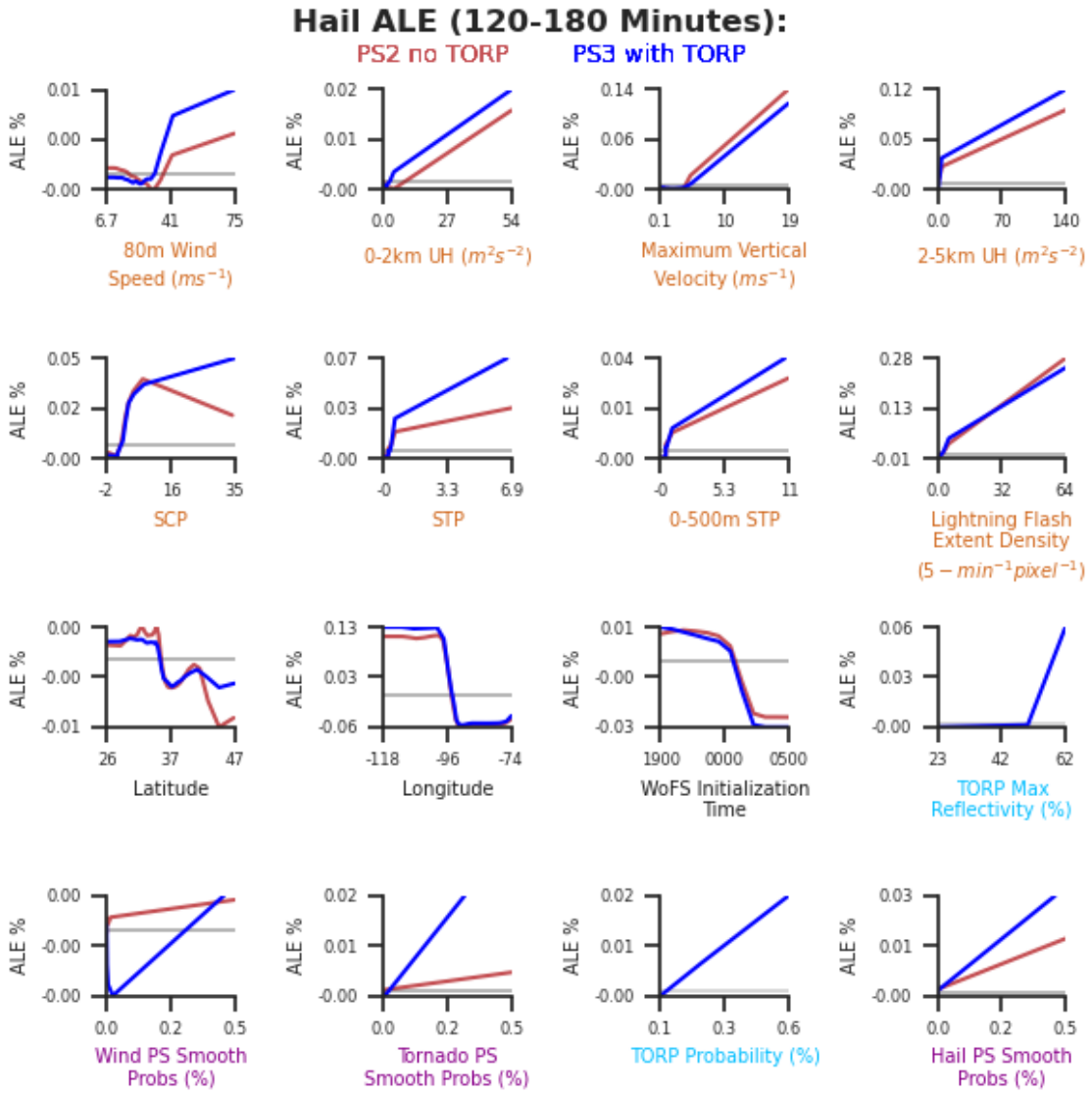

Figure 4.15: As in Fig. 4.13 but for the 120-180 minute period

4.2.1.2 Wind

Similar to hail, most WoFS predictors do not change much in influence between the two models in the 0-60 minute period, but SCP does increase to have modest influence when PS3 and TORP are used (from nearly no influence when PS2 and no TORP is used; Fig. 4.16). An even bigger change can be seen with 80 m wind speed.

Despite being known to be correlated to severe wind (Kucher and D. 2006), its influence drops approximately in half with the change in the predictor dataset. It is also interesting that there is a clear threshold at approximately 40 ms^{-1} at which 80 m wind speed becomes more influential (Fig. 4.16).

For ProbSevere predictors, there is no change to the influence of wind ProbSevere probabilities between PS2 and PS3 (Fig. 4.16). Tornado and hail ProbSevere probability influence decreases from PS2 to PS3 (Fig. 4.16). This is consistent with the independence of wind forecasts from hail and tornado forecasts shown in the previous section. While less influential than for hail (as expected), TORP maximum reflectivity still is modestly influential to wind forecasts, as is TORP probability (Fig. 4.16).

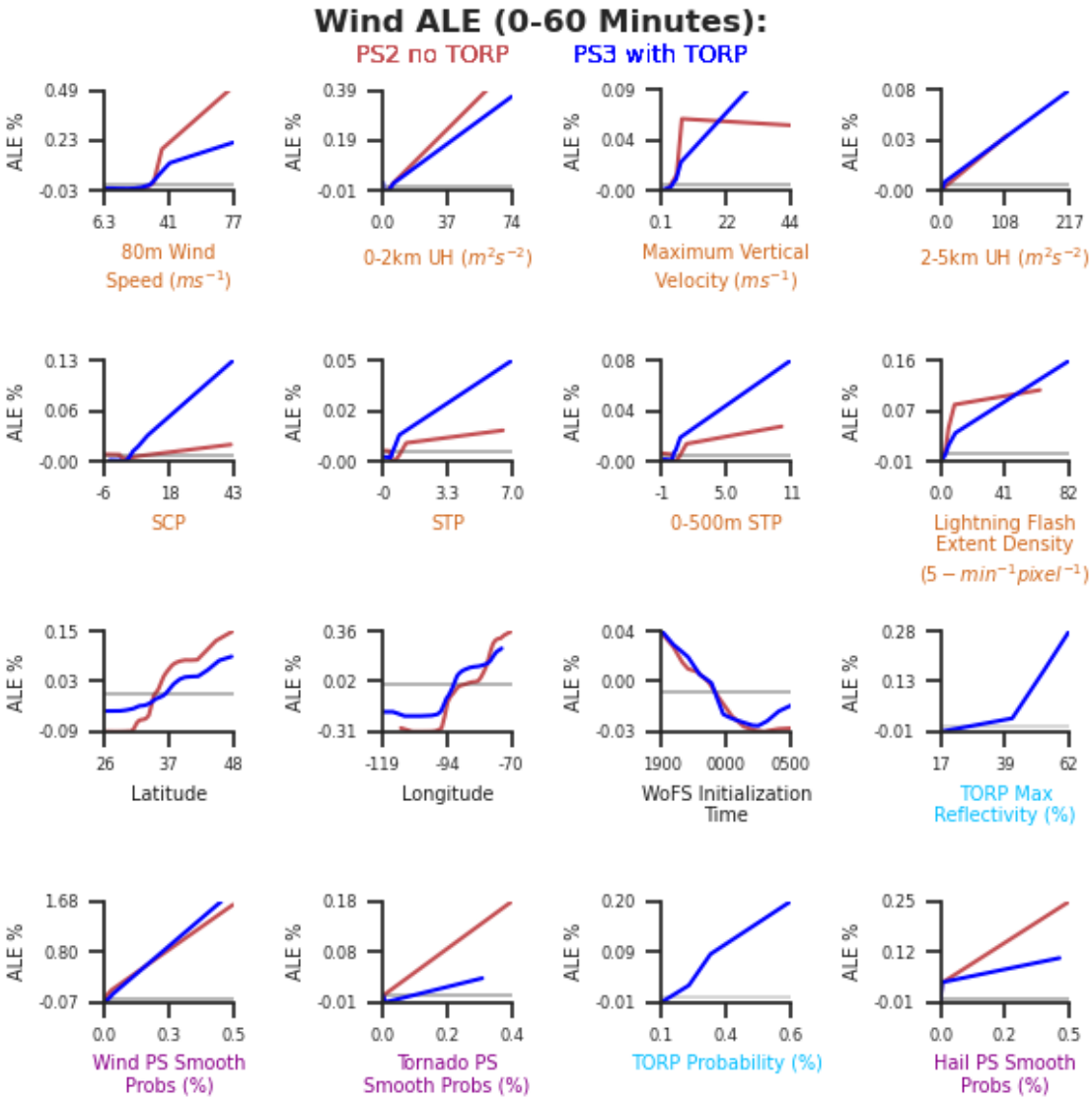

Figure 4.16: As in Fig. 4.13, but for wind.

In the 60-120 minute period, there are no noteworthy changes in predictor influence between the two models. However, there is a stronger pattern of increased WoFS influence compared to what was seen with hail between the 0-60 and 60-120 minute periods. Seven of the eight WoFS predictors shown increase in influence in the 60-120 minute period compared to the 0-60 minute period, and the last (80 m wind speed) does not change (decreases) in importance for the PS3 and TORP (PS2

without TORP) model (Fig. 4.17). While there is an influence from longitude, and to an extent latitude, in the 0-60 minute period (higher forecasts to the northeast, lower forecasts to the southwest), there is a stronger longitudinal gradient in influence with -96° becoming an important cutoff once again. While it seems to be a more obvious impact of overfitting with the stronger influence shown with wind LSRs, it is odd that the northeast is favored for wind LSRs over any other region instead of the southeast as would be expected (Smith et al. 2013; Edwards et al. 2018). It does appear that the eastern bias is more influential than the northern bias, so it's possible that this bias is more generalized across the eastern US and even larger in the northeast but that the focus of bias studies have been on the southeastern (compared to the northeastern) US due to the increased climatology of severe events. It is also possible that there is overfitting to an incorrect pattern, especially with few WoFS domains over the eastern US in general for WoFS-PHI to train on (Fig. 2.4).

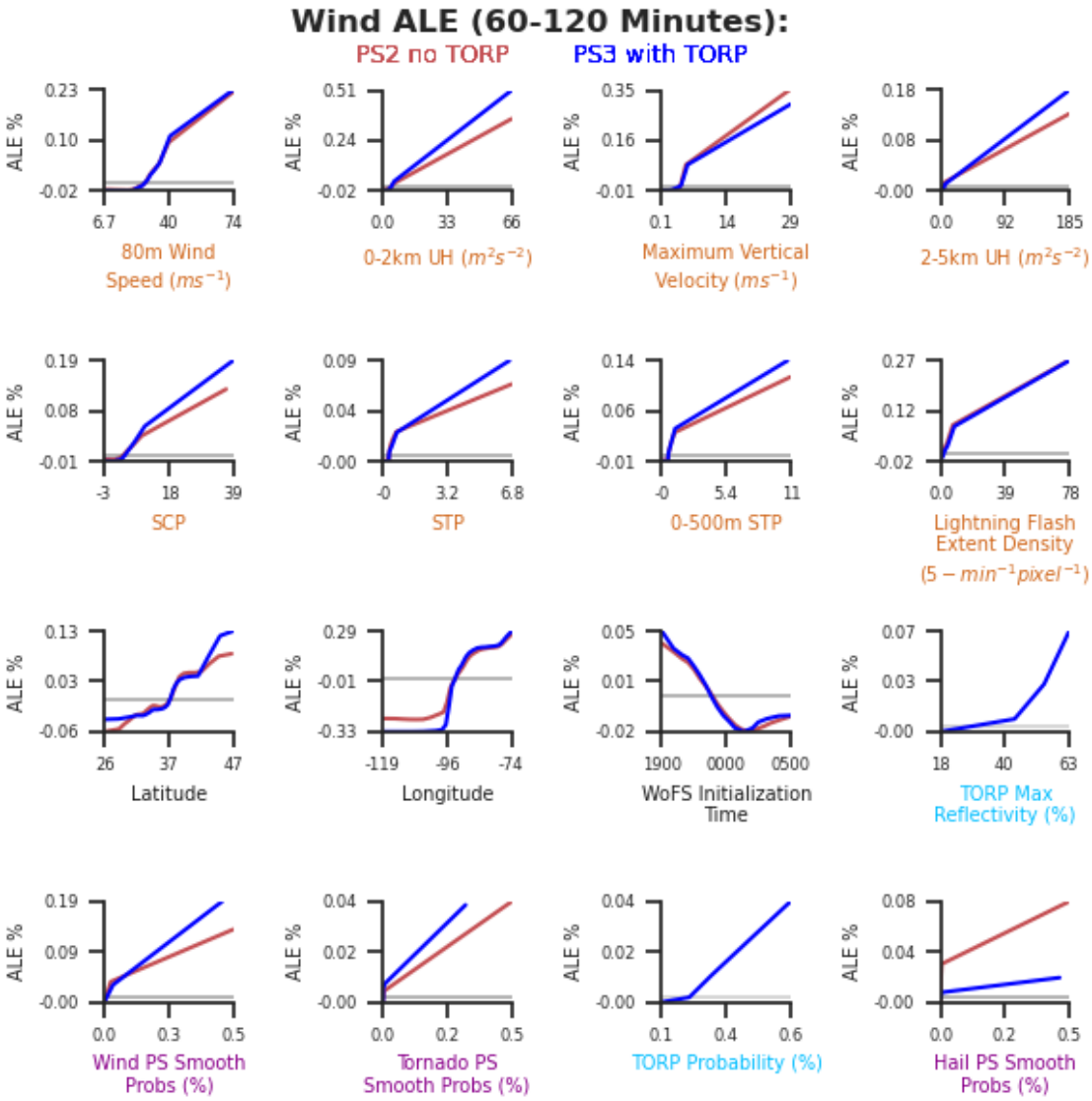

Figure 4.17: As in Fig. 4.16 but for the 60-120 minute period

In the 120-180 minute period, most predictors decrease in influence relative to the 60-120 minute period, similar to with hail and perhaps reflective of decreasing sharpness (Fig. 4.18). It is also interesting that the influence of latitude and (especially) longitude decrease in influence with lead time. It would be expected that as uncertainty increases and meteorological predictors become less correlated to

events, predictors would be used more evenly, and overfitting to predictors like latitude and longitude would increase.

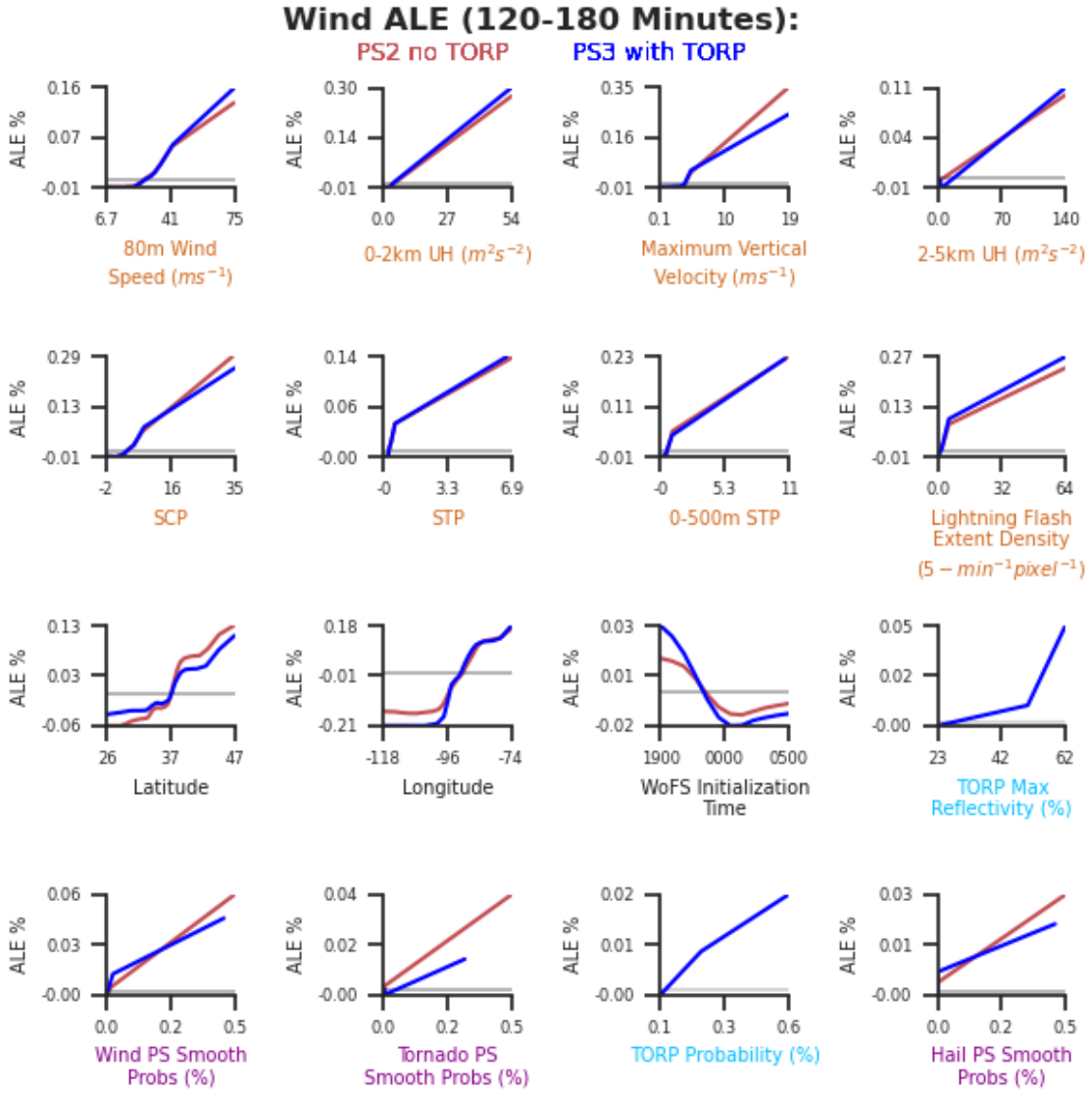

Figure 4.18: As in Fig. 4.16 but for the 120-180 minute period

4.2.1.3 Tornadoes

Tornado forecasts have much lower ALE values in general compared to hail and wind forecasts (Fig. 4.19), reflective of their much lower sharpness. There is minimal

change in any WoFS predictors between the PS2 without TORP and PS3 with TORP models, and most WoFS predictors are hardly influential at all in the 0-60 minute period aside from STP and SCP (Fig. 4.19). For ProbSevere probabilities, PS3 tornado probabilities are expectedly more influential than PS2 tornado probabilities. Hail ProbSevere probabilities have very low influences regardless of Probsevere version, and PS3 wind probabilities are much less influential than PS2 wind probabilities (Fig. 4.19). This decrease in the influence of wind ProbSevere probabilities is consistent with the decrease in the influence of tornado ProbSevere probabilities for wind forecasts (Fig. 4.16). Both of these patterns are evidence that PS3 allows WoFS-PHI to discriminate between hazards better than PS2.

The influence of TORP probabilities is enlightening. There is a clear influence, but only when it is larger than approximately 40%. This suggests that only higher-end TORP objects are influential to WoFS-PHI, and this may be reflective of TORP's known overforecasting bias (Sandmael et al. 2023).

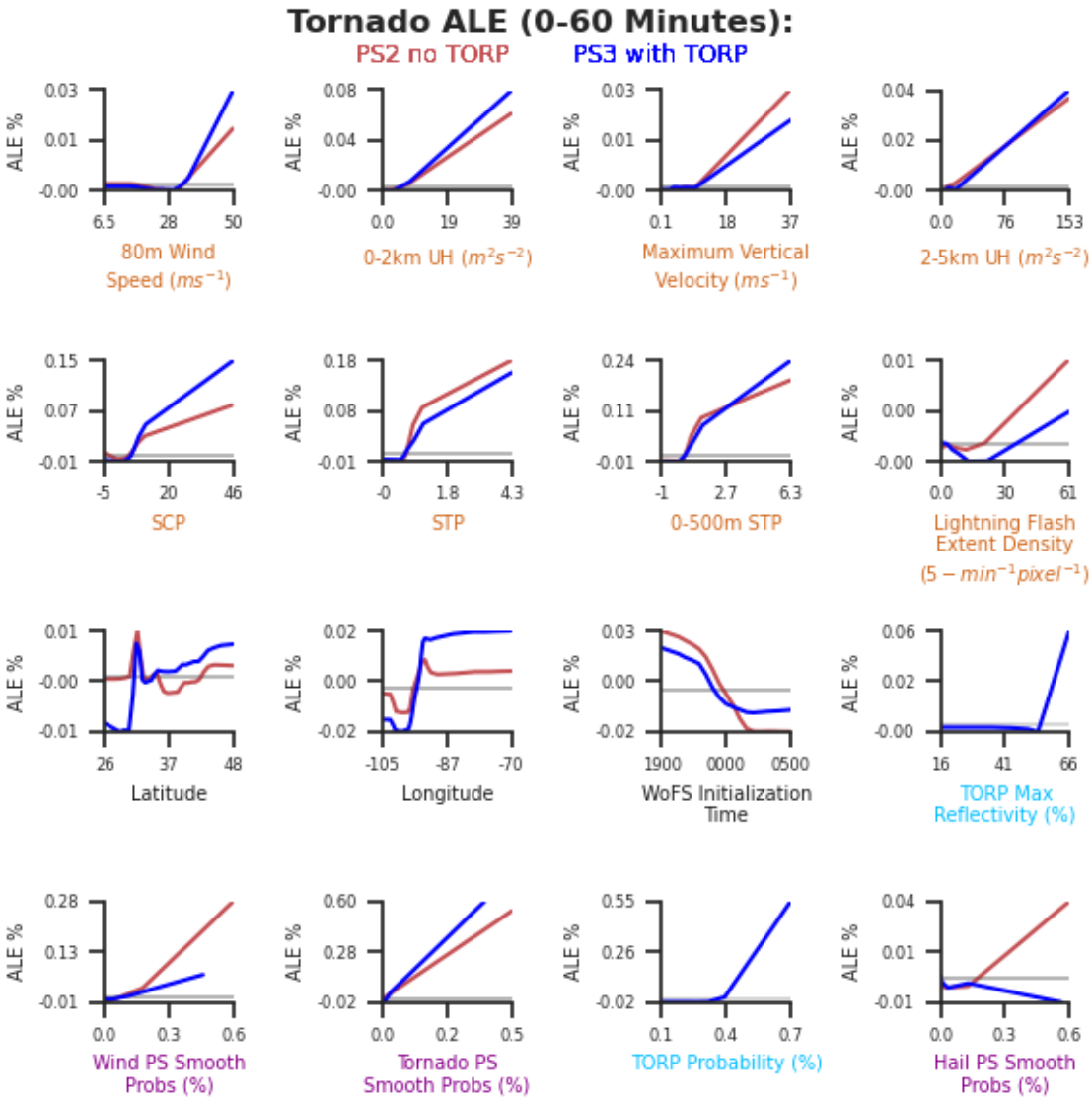

Figure 4.19: As in Fig. 4.13, but for tornadoes.

In the 60-120 minute period, there are negligible changes to influence among WoFS predictors between models and from the 0-60 minute period (Fig. 4.20). The only noteworthy difference in predictor influence between models that remains is that PS3 tornado probabilities are more influential than PS2 tornado probabilities, although even PS3 tornado probabilities are less influential than in the 0-60 minute period (Fig. 4.20). The effect of lead time on TORP influence is even more dramatic,

as there is effectively no TORP influence in the 60-120 minute period. There is also very little change in the 120-180 minute period aside from a continued lack of influence from ProbSevere tornado probabilites, although PS3 is still more influential than PS2 (Fig. 4.21).

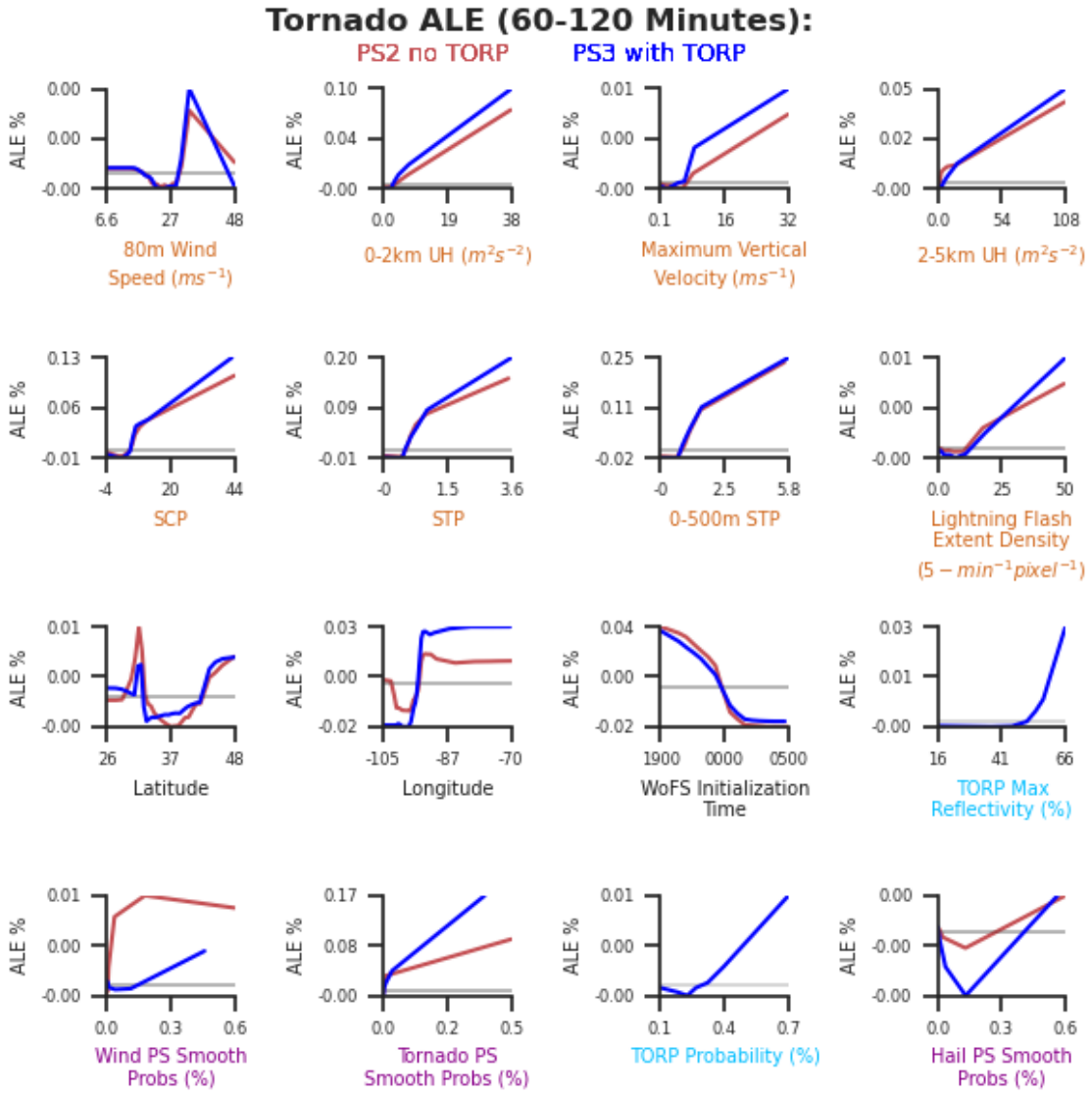

Figure 4.20: As in Fig. 4.19 but for the 60-120 minute period

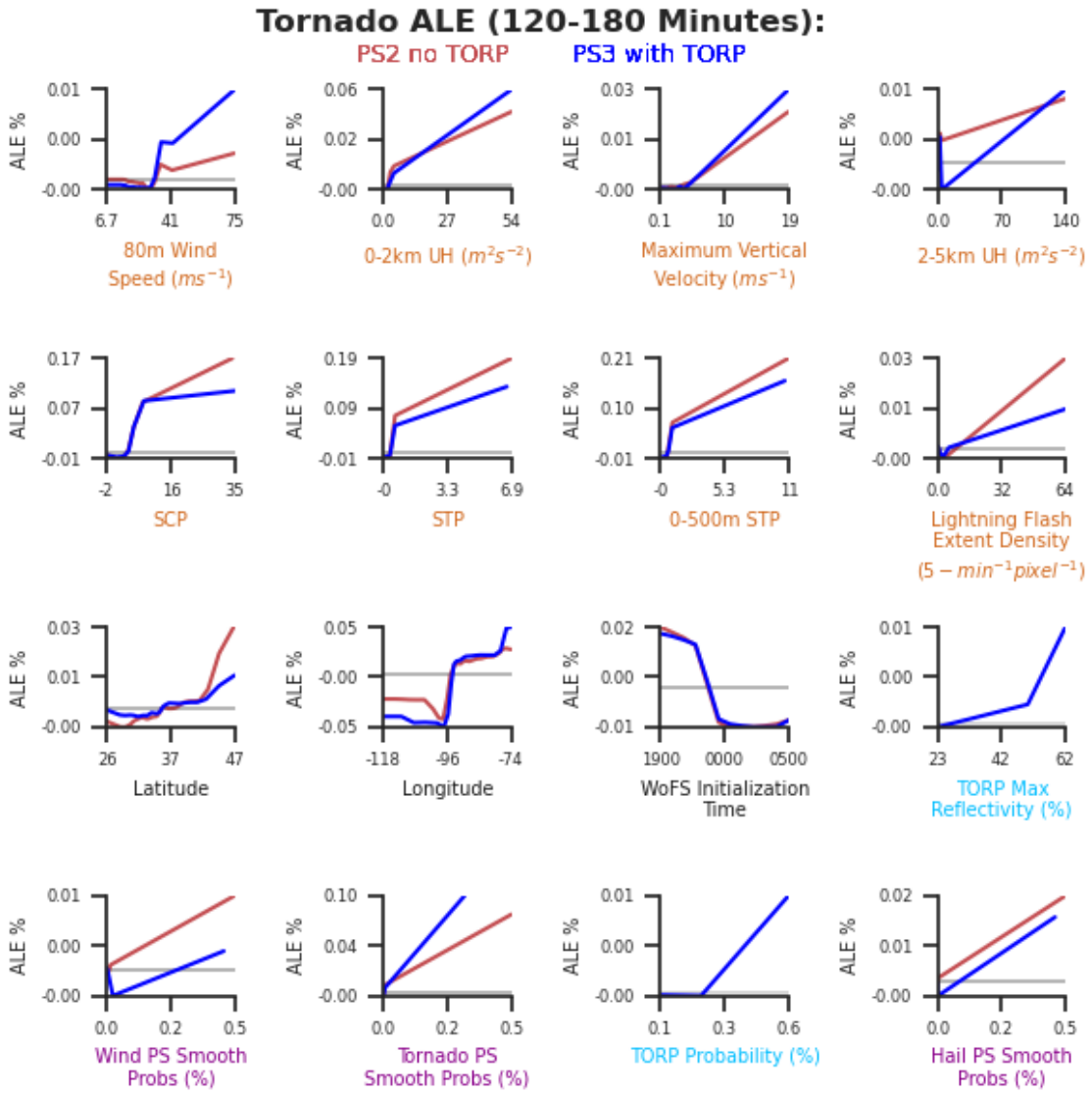

Figure 4.21: As in Fig. 4.19 but for the 120-180 minute period

4.2.2 Impact of TORP without PS3

4.2.2.1 Hail

Expectedly, hail forecasts are increasingly influenced by the same three WoFS predictors as with previous models in the 0-60 minute period (updraft speed, $2-5$ km UH, and lightning flash extent density; Fig. 4.22). This makes sense since it has

already been shown that WoFS predictors are far more important than TORP for the WoFS and TORP only model (Fig. 4.10). However, despite its lack of overall importance, TORP predictors are quite influential when they are available, especially TORP maximum reflectivity (Fig. 4.22). In fact, the influence of high-end reflectivity on hail forecasts when PS3 is unavailable far exceeds any other ALE magnitude seen for any other predictor in any other forecast across all hazards, models, target datasets, and lead times (with the caveat that comparing ALE magnitude across different predictors is not always fair). This suggests that TORP is actually a valuable predictor, but that its lack of availability is the main impediment on it being a more important overall predictor. These results suggest that if there was a similar product that identified the maximum reflectivity (and perhaps other radar fields known to be correlated with severe hail) for all storms regardless of tornadic potential, that could be a very important predictor group for WoFS-PHI. Additionally, even the TORP probability has an outsized impact on hail forecasts in the early lead times, though not nearly as large as the maximum reflectivity (Fig. 4.22). This result is similar to what was seen with the comparison between PS2 and PS3 where the improved PS3 tornado probabilities were more influential to hail forecasts than with PS2 (Fig. 4.13). The suggestion that hail and tornado forecasts may be decently well correlated seems to be highlighted again here. This could be due to the fact that strong tornadic supercells would also be likely to be large hail producers. The longitudinal influence is also stronger when PS3 is not used in the predictors (Fig. 4.22).

A simple analysis of grid points covered by TORP/PS3 objects supports the claim that TORP's biggest weakness is a lack of availability. Generally, only TORP objects with a TORP probability above 20% have any influence for hail forecasts. While about 1.02% of all gridpoints are associated with a TORP object, only about 0.49%

are associated with TORP objects with a probability greater than 20%. For reference, about 1.49% have non-zero PS3 hail probabilities. It is also important to recall that while only TORP objects with a probability above approximately 20% are influential to hail forecasts, all PS3 hail probabilities are influential. So, while TORP objects are only about 1.5 times less common than PS3 objects, *influential* TORP objects are about 3 times less common than *influential* PS3 objects.

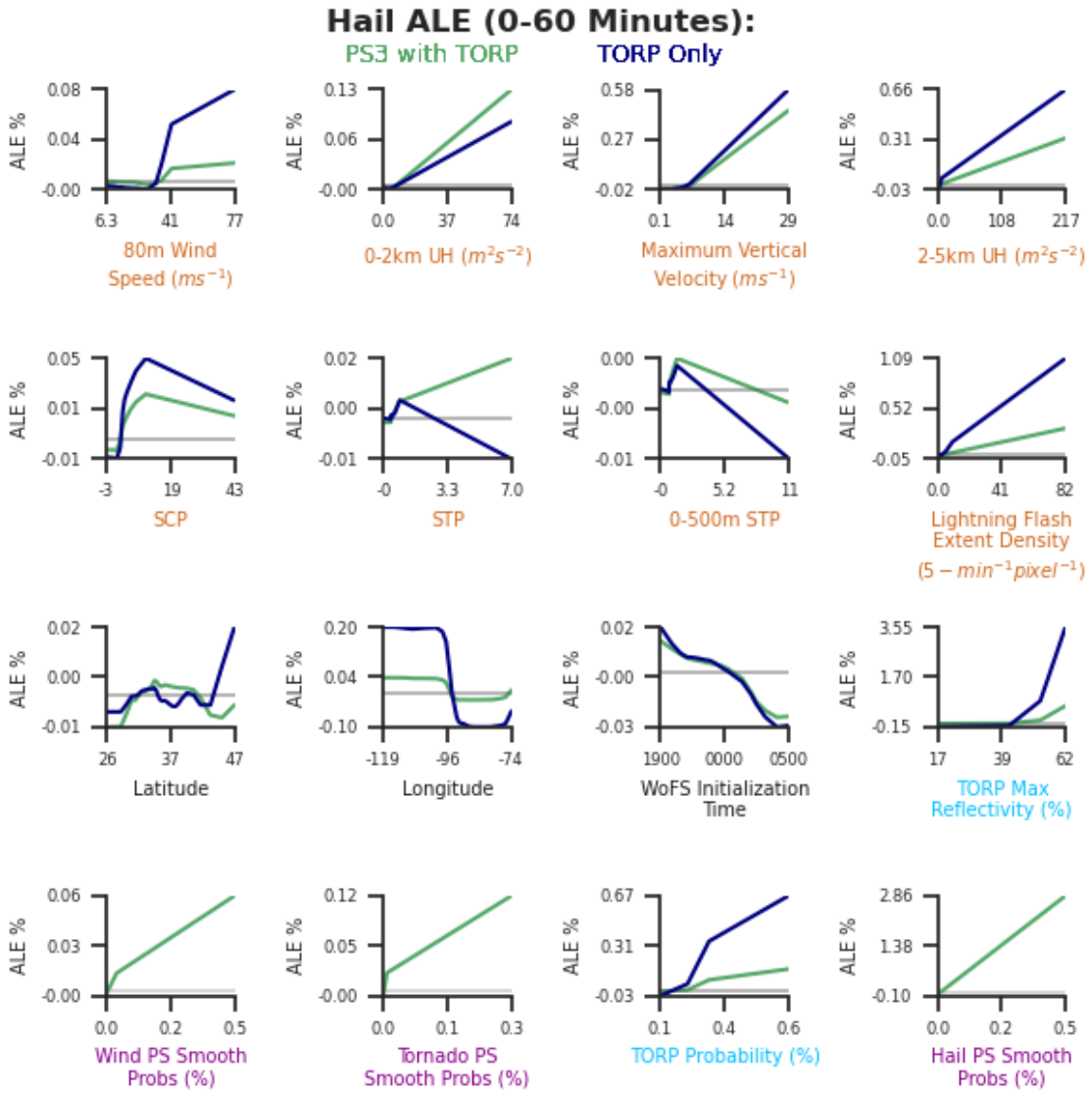

Figure 4.22: As in Fig. 4.13, but for a comparison between the model trained with PS3 and TORP (green) and TORP only (navy). Both models are trained against LSRs and warnings.

Even in the 60-120 minute period, WoFS-PHI retains influence from the maximum reflectivity predictor that is not seen when PS3 is used (Fig. 4.23). Most WoFS predictors do not change as much in influence relative to the model that includes PS3 at this later lead time (Fig. 4.23), and this is reflected in the decreasing difference in performance metrics at later lead times. By the 120-180 minute period,

TORP predictors are no longer particularly influential, and there is minimal difference in the influence of WoFS predictors between the model with and without PS3 (Fig. 4.24). Longitude retains its influence at the later lead times as well, and is most influential in the 60-120 minute period (Figs. 4.23, 4.24).

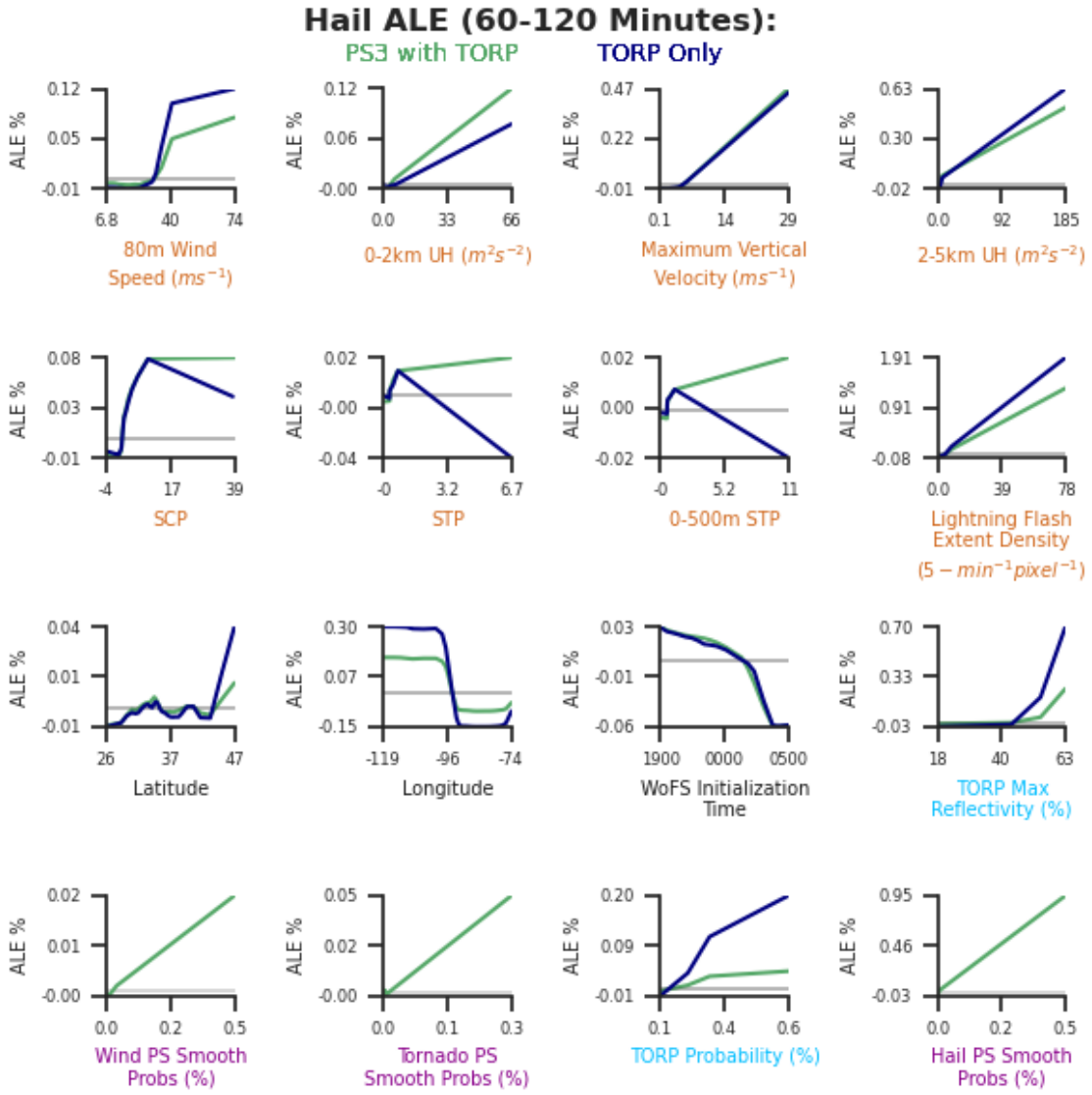

Figure 4.23: As in Fig. 4.22 but for the 60-120 minute period

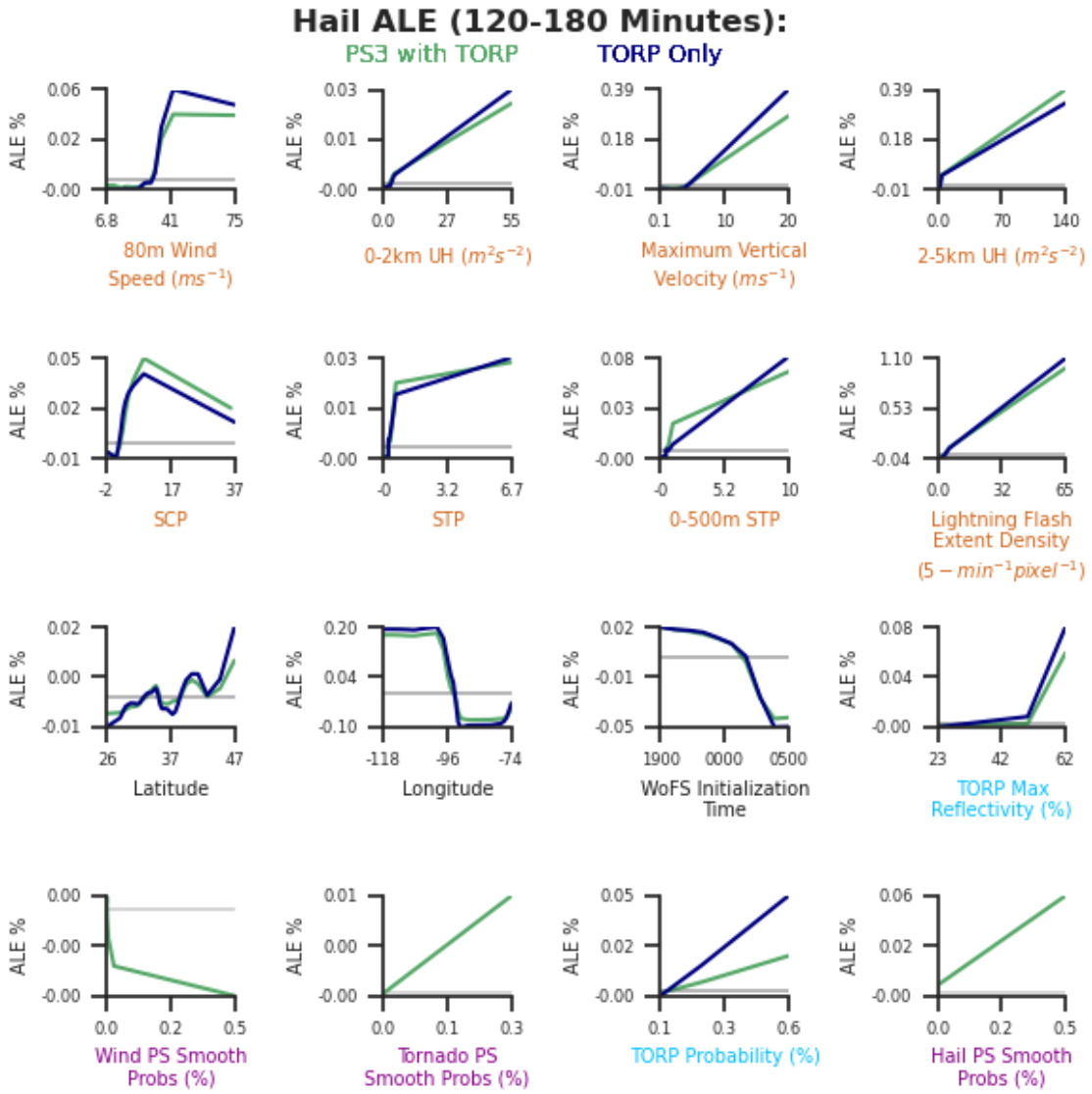

Figure 4.24: As in Fig. 4.22 but for the 120-180 minute period

4.2.2.2 Wind

Similar to hail, the same WoFS predictors (80 m wind speed, updraft speed, $2-5$ km UH, and lightning flash extent density) continue to increase in influence in the absence of PS3 as in previous wind forecast experiments (Fig. 4.25). In addition, SCP also increases in influence when PS3 is removed (Fig. 4.25). There is also similar

influence from TORP probability as with hail forecasts, and influence from maximum reflectivity, although not nearly as strong as with hail forecasts (Fig. 4.25). For both TORP predictors, their influences are much greater in the absence of PS3. Once again, it appears as though TORP heavily influences the final RF probabilities, but it is likely limited in terms of an overall impact by the lack of TORP objects to use. A similar grid point analysis as done for hail reveals that about 2.16% of all gridpoints are associated with non-zero PS3 wind probabilities. Thus, influential TORP objects appear to be more than 4 times less common than influential PS3 objects for wind forecasts.

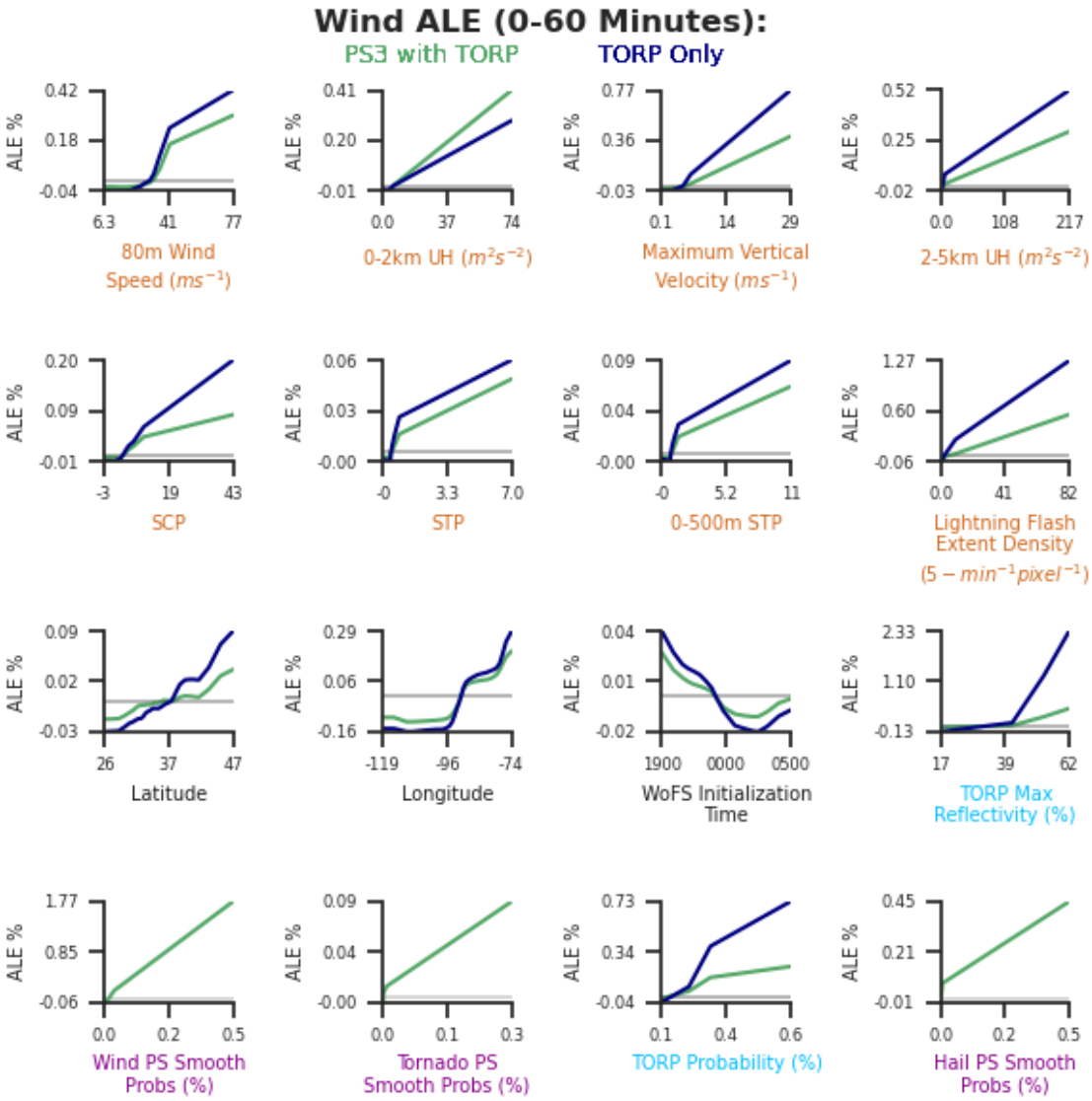

Figure 4.25: As in Fig. 4.22, but for wind.

In the 60-120 minute period, WoFS predictor influence is relatively unchanged with the removal of PS3, although there is slightly more influence from the lightning flash extent density (Fig. 4.26). Less change compared to the model with PS3 would be expected relative to the 0-60 minute period since PS3 was less impactful at later lead times to begin with. There is also still modest contributions from TORP predictors, especially maximum reflectivity, but it is less than in the 0-60 minute

period and compared to hail in the 60-120 minute period (Fig. 4.25). By the 120-180 minute period, the WoFS predictor influences are nearly identical to the model with PS3 (Fig. 4.27), which is expected given nearly identical performance at the latest lead times. Additionally, TORP probability has dropped to having nearly no influence, and TORP maximum reflectivity has only a slight influence (Fig. 4.27), which is also expected given that TORP is based on current observations, and it's predictive capabilities diminish quickly with lead time.

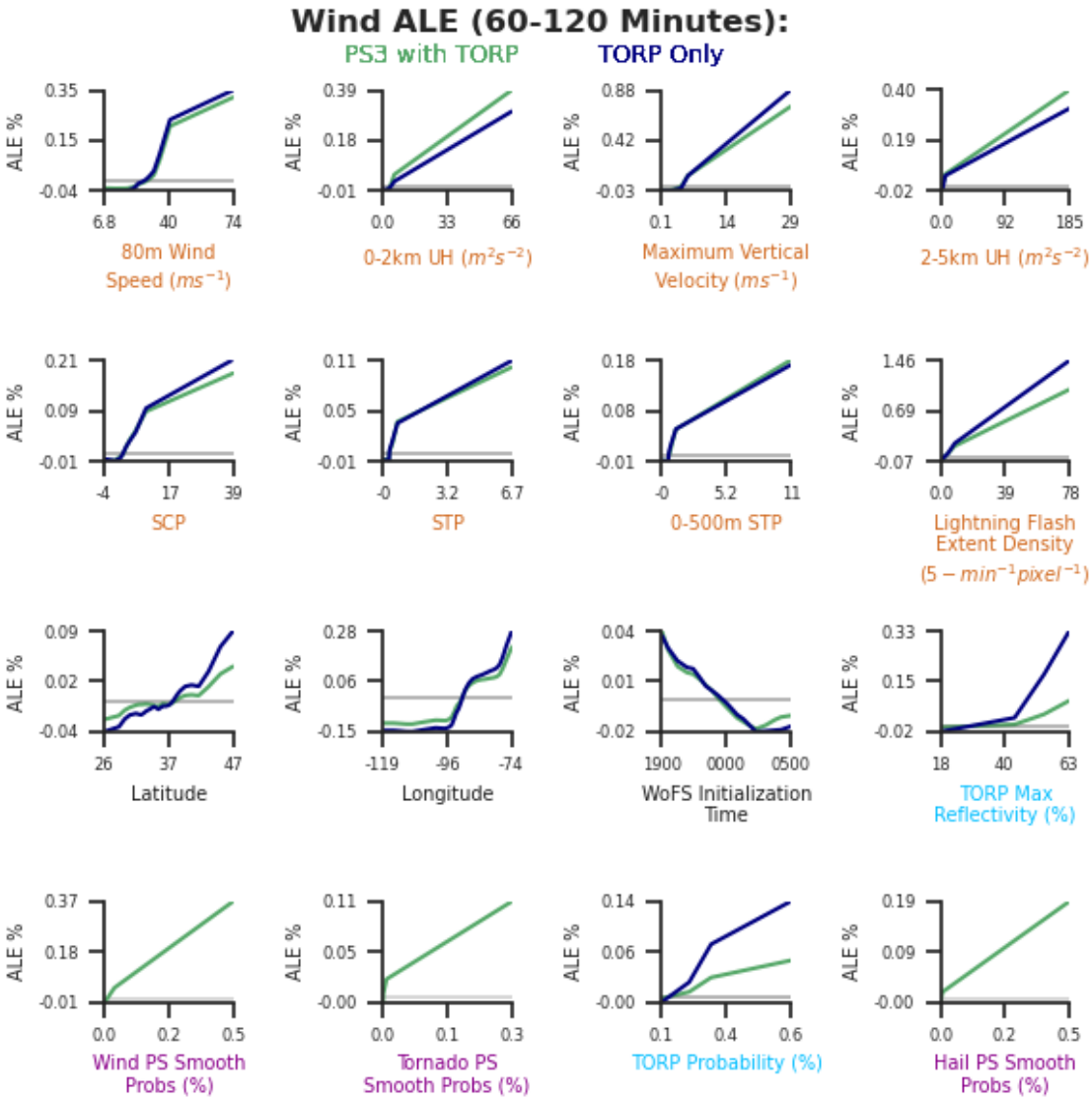

Figure 4.26: As in Fig. 4.25 but for the 60-120 minute period

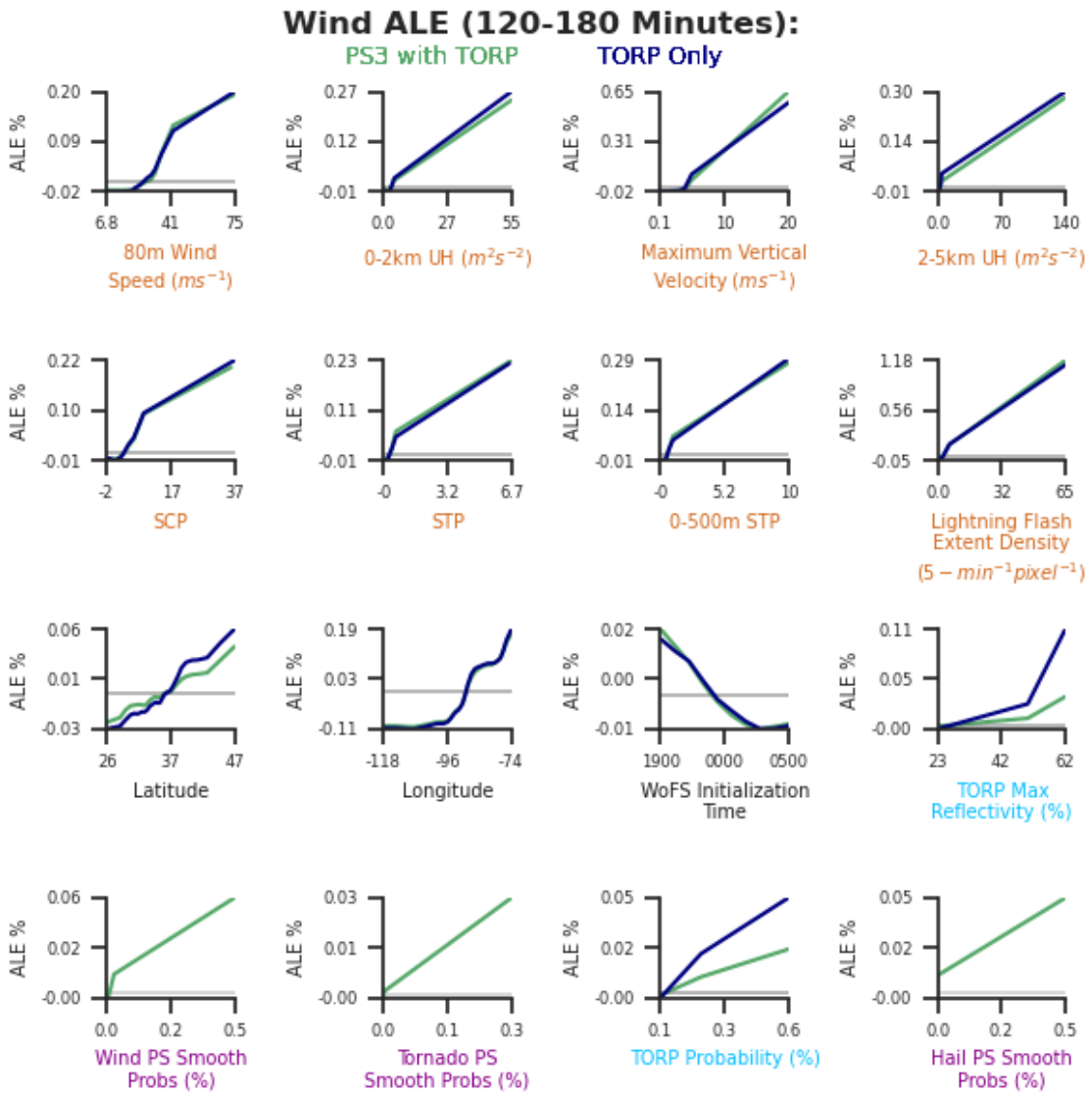

Figure 4.27: As in Fig. 4.25 but for the 120-180 minute period

4.2.2.3 Tornadoes

Interestingly, not all of the typical WoFS tornado predictors ($0-2$ km UH, SCP, both forms of STP) increase in influence in the absence of PS3 (Fig. 4.28). In fact, of those four predictors, only SCP increases in influence with the removal of PS3 data, and the other four decrease in influence. WoFS $2-5$ km UH increases in importance,

and this could make up for the decreased influence from other predictors when this effect is spread across all $2-5$ km UH predictors.

TORP predictor behavior is also interesting. Both hail and wind forecasts saw influence from TORP maximum reflectivity, and still large influence from TORP probability (Figs. 4.22, 4.25). While the influence of TORP probability is larger than for hail or wind, it is still much smaller than the influence of maximum reflectivity on hail and wind forecasts (Fig. 4.28). It was shown that TORP had the least impact for tornado forecast performance, and the fact that the most influential TORP predictor is more influential for hail and wind forecasts than tornado forecasts. This could be due to the nature of tornado forecasts being more uncertain leading to lower sharpness than hail or wind forecasts. TORP probability was nearly exactly as influential as tornado PS3 probability (Fig. 4.28). The reason for TORP not being as important as PS3 appears to again be due to the very low number of TORP objects rather than the lack of influence when TORP objects did exist. The grid point analysis of PS3 tornado objects reveals that approximately 1.30% of grid points were associated with non-zero PS3 tornado probabilities. However, TORP objects were not influential to tornado forecasts until about 30% (40% when PS3 was included) rather than the 20% threshold for hail and wind forecasts. Only about 0.20% of grid points were associated with influential TORP objects when PS3 is not included. When including PS3 and using the 40% threshold, only about 0.08% of grid points are covered by influential TORP objects. Thus, TORP objects influential to tornado forecasts are 6-13 times less common than influential PS3 objects, a larger dropoff than for wind or hail.

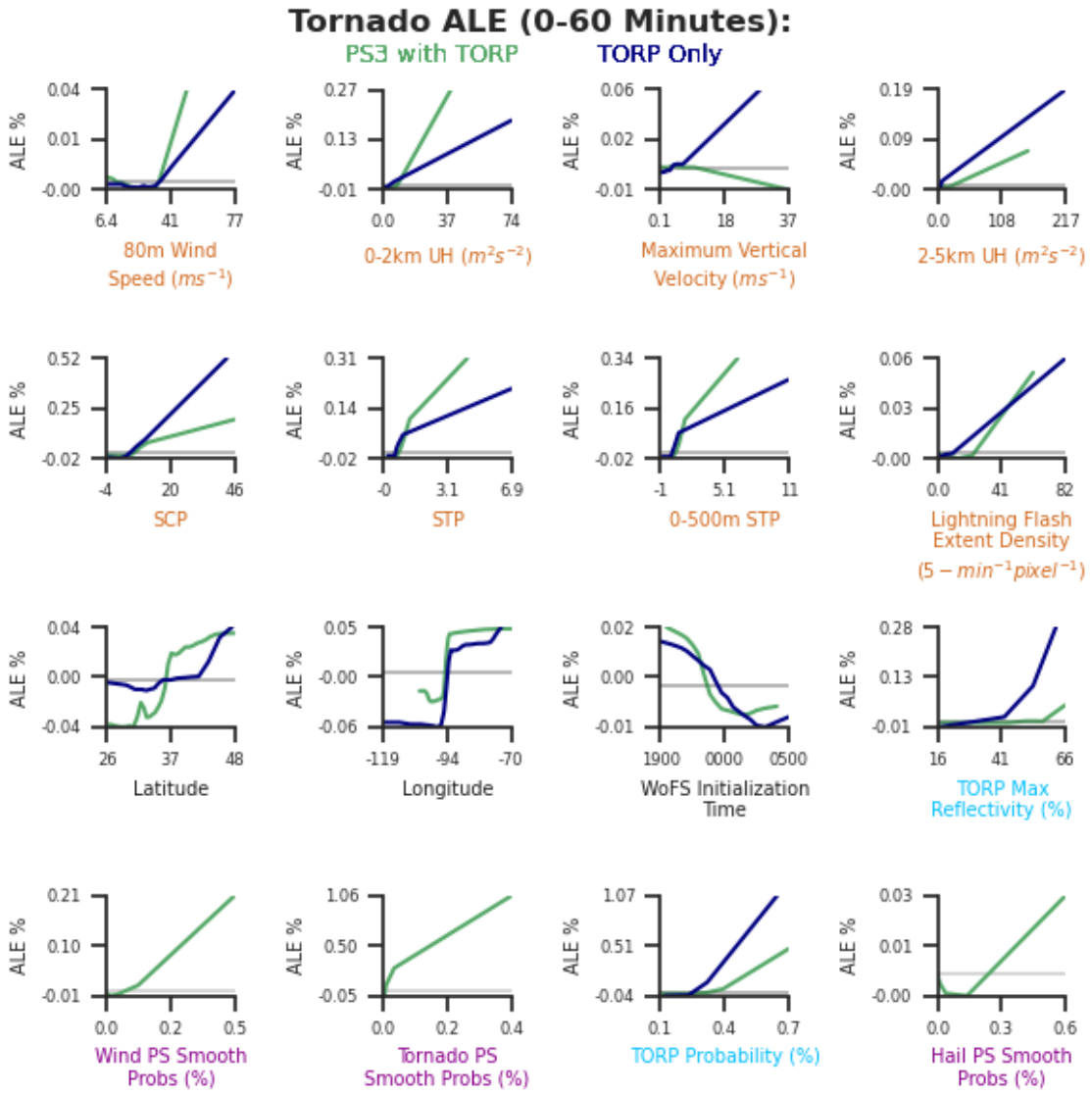

Figure 4.28: As in Fig. 4.22, but for tornadoes.

WoFS predictor behavior relative to the model that includes PS3 is very similar in the 60-120 minute period (Fig. 4.29). Similarly most influential WoFS predictors ($0-2$ km UH, both forms of STP) are once again less influential than when PS3 is included with the exception again being SCP(Fig. 4.29). TORP has a small influence during the 60-120 minute period (ALE not reaching even 0.05% even for high-end TORP objects; Fig. 4.29). The fact that TORP retains influence to the 60-120

minute period for hail and wind but not tornadoes is likely due to the short-lived nature of tornadoes compared to hail and wind events that can last longer more often. In the 120-180 minute period, TORP predictors are reduced to a negligible influence (similar to with hail and wind forecasts; Fig. 4.30). Simultaneously, all WoFS predictors with non-negligible influence are equally or more influential when PS3 is removed, a deviation from the pattern at earlier lead times. There is no increased sharpness with the non-PS3 model, though (Fig. 3.20). This is likely due to the influence of PS3 that is not replicated by TORP probability when PS3 is removed (Fig. 4.30). Thus, while WoFS predictors are contributing to increased sharpness, this increase is being cancelled by essentially a total loss of the PS3 influence.

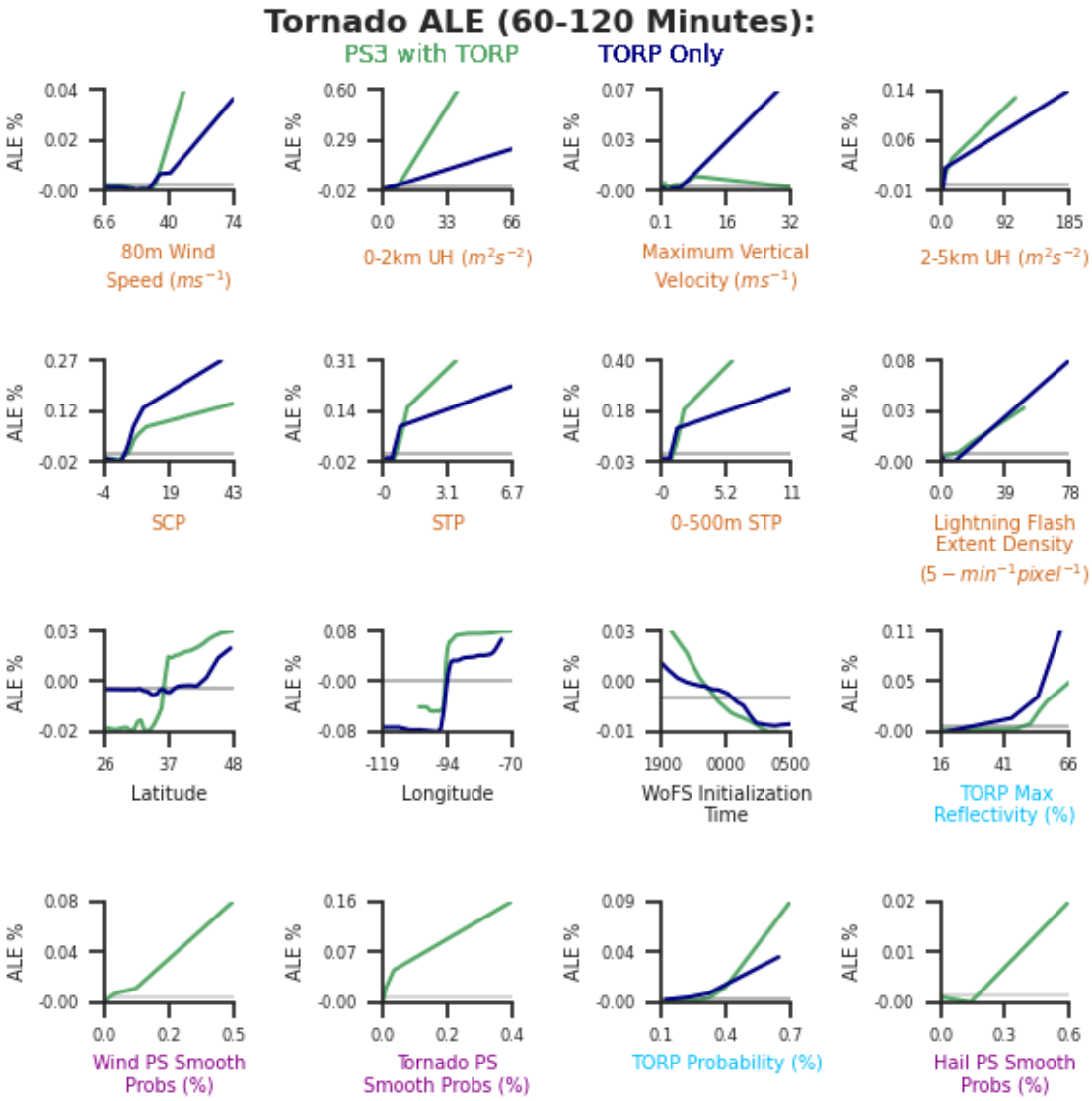

Figure 4.29: As in Fig. 4.28 but for the 60-120 minute period

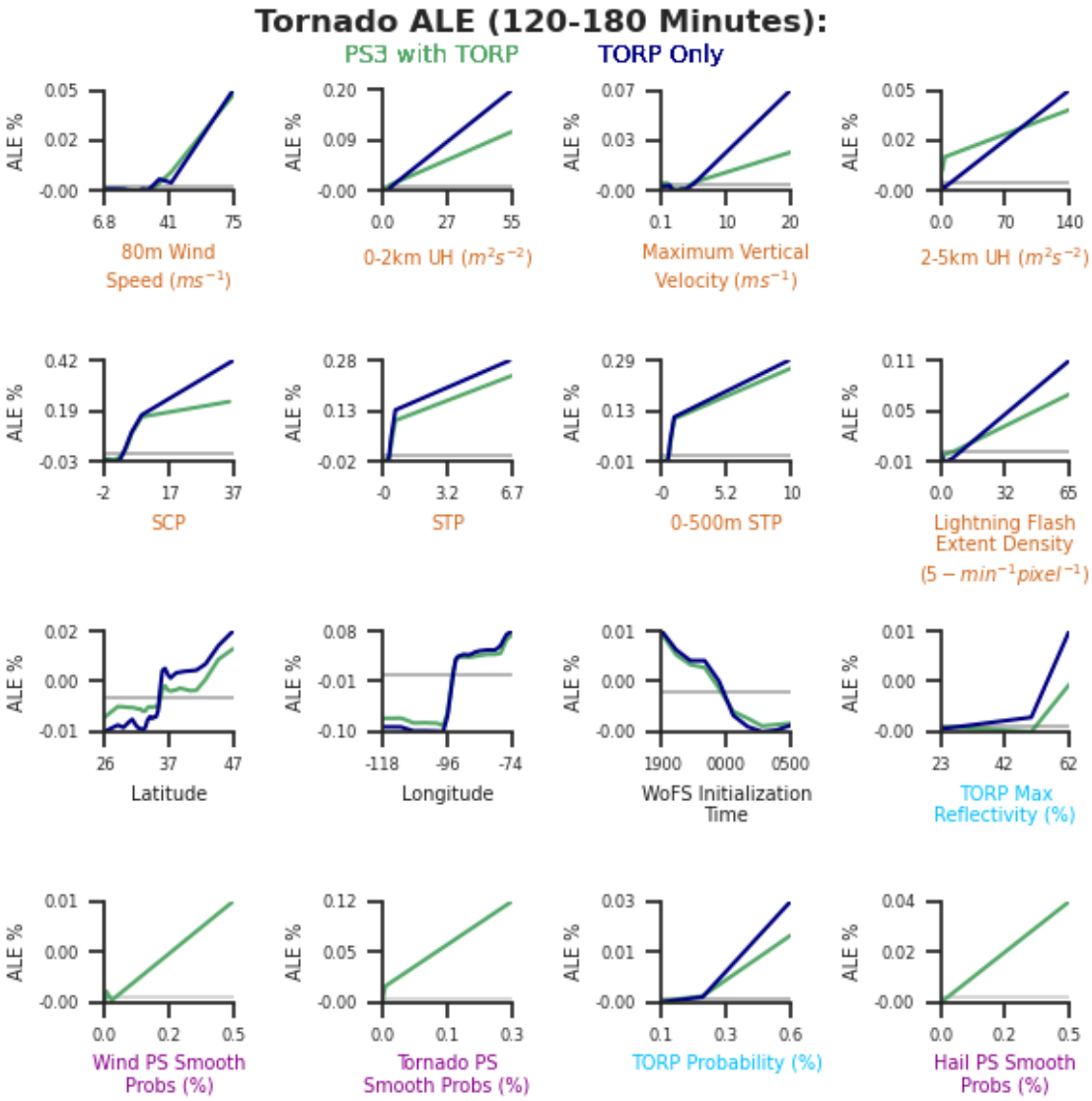

Figure 4.30: As in Fig. 4.28 but for the 120-180 minute period

Chapter 5

Case Studies

Two example forecast initializations will be used as case studies to highlight example behavior of WoFS-PHI using different datasets and event types. Example forecasts show the effect of including PS3 and TORP as predictors, the impact of using warnings as targets during in training, and the impact of using TORP as the only observational predictor dataset. Lastly, a comparison between the baseline (PS2, no TORP, trained on LSRs-only) and the most skillful (PS3, with TORP, trained on LSRs and warnings) models is shown to highlight the total effect of all improvements made.

The two testing dataset forecast initializations used will be April 30, 2019 at 0100z (8:00pm local time) and April 28, 2020 at 2100z (4:00pm local time). These forecasts are chosen because they represent different times of day, and there is at least one LSR or warning for each hazard at each lead time.

5.1 April 30, 2019

5.1.1 Impact of Changing the Predictor Dataset

For the forecast initialized on April 30, 2019 at 0100z, the baseline model clearly highlights wind as the most likely hazard, and it correctly places the highest wind probabilities within the cluster of wind LSRs in southwestern MO (Fig. 5.1g-l). The

wind forecasts lose spatial precision with lead time, though there are relatively high magnitudes maintained out to 150 minutes of lead time. This is not surprising given that wind was shown to have the least decrease in performance with lead time. There are hardly any hail probabilities issued at any lead time (Fig. 5.1m-r), and tornado probabilities are generally low, though the southern area (south-central OK) is highlighted slightly more than the northern area (southwestern MO, Fig. 5.1a-f). The tornado forecasts also successfully highlight the dual threat areas rather than one spatially continuous area.

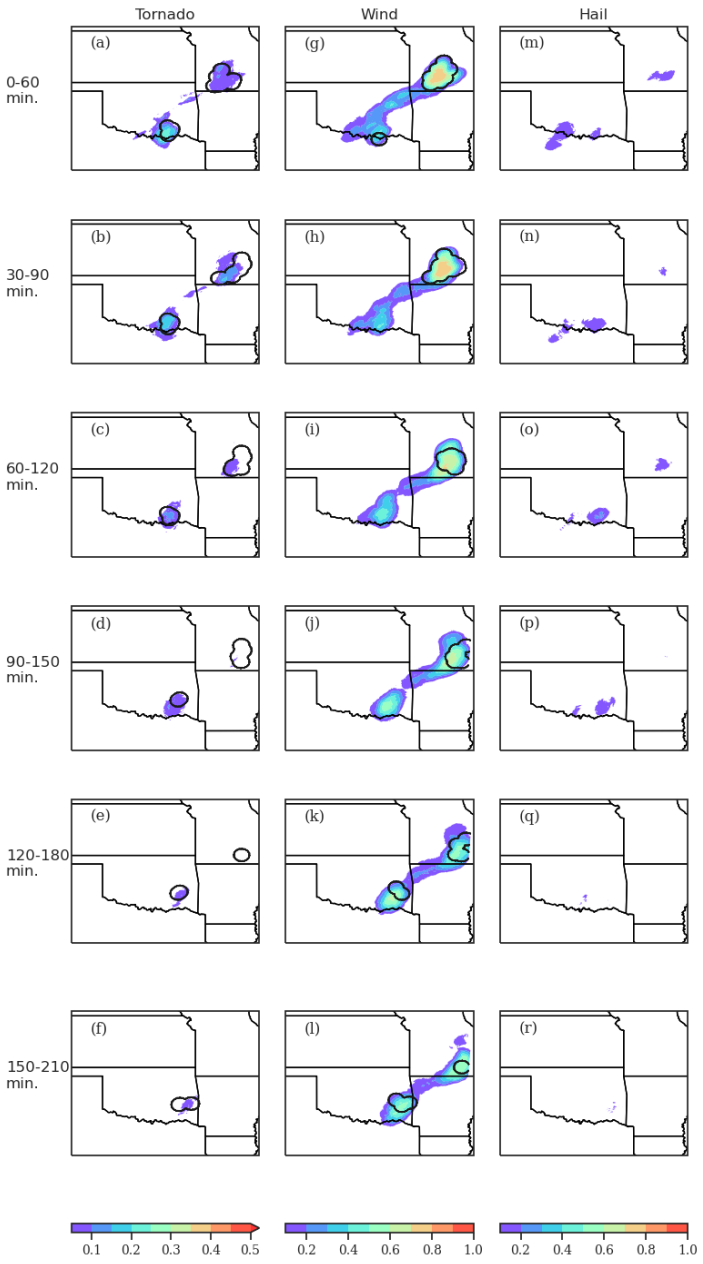

Figure 5.1: The forecasted probabilities from WoFS-PHI using PS2 and no TORP and trained on LSRs-only (i.e., the baseline model configuration) for April 30, 2019 from 8:30pm to 12am local time. The filled contours represent the WoFS-PHI probabilities, and the black contours represent the 39 km radii around hazard-specific LSRs.

When PS3 and TORP are used, there is nearly no change in hail forecasts (Fig. 5.2m-r), and this is due to the nearly non-existent hail probabilities in either forecast. The main difference in wind probabilities appears to be lowered forecast probabilities at early lead times in central OK (Fig. 5.2g-i), however, these lowered probabilities with PS3 and TORP data occur in areas that have no wind LSRs, so this new data can be seen still positively influencing the forecasts, just in terms of correct negatives rather than hits. There is a clear signal of increased tornado probabilities in southwestern MO in the 0-60 minute forecast period (Fig. 5.2a), and this is spatially consistent with the largest cluster of tornado LSRs. The fact that tornado forecasts were most impacted by the new data in this forecast is unsurprising as the tornado models showed the biggest improvement in forecast metrics after the addition of new data, especially in the 0-60 minute period (Fig. 3.3c). There is nearly no influence from the new datasets for 90 minute lead times and beyond for any hazard (Fig. 5.2d-f, j-l, p-r). This is also unsurprising as the influence of PS3 and TORP data decreases with lead time, and the improvement in model performance associated with including this data similarly decreases with lead time.

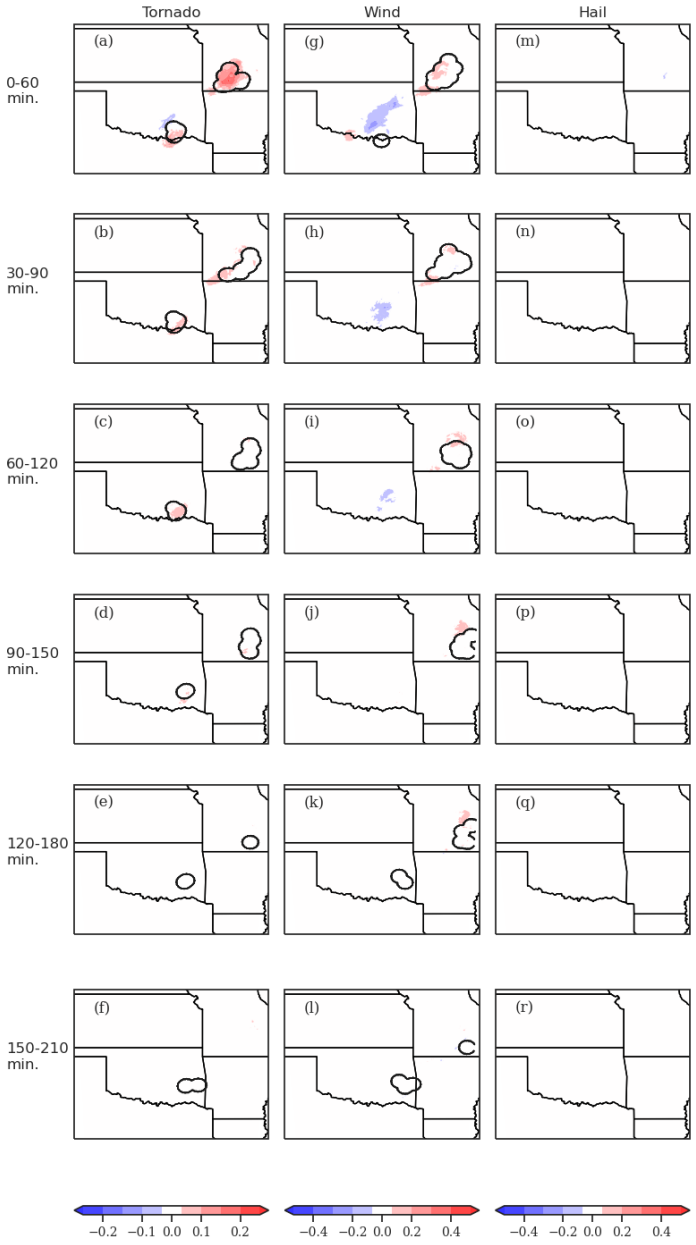

Figure 5.2: The difference between the probabilities forecasted by the PS3 with TORP model and PS2 no TORP model. Red represents areas where the PS3 with TORP model had higher probabilities, and blue represents areas where the PS2 no TORP model had higher probabilities. For each of these maps, the model with lesser forecast performance is subtracted from the model with greater forecast performance

5.1.2 Impact of Training on Warnings

The most immediately obvious impact from training on warnings for the April 30, 2019 case is that of substantial increases in tornado probabilities (Fig. 5.3a-f). While this increase in probabilities is greatest at early lead times, the signal extends to all lead times. The impact is also greatest in the southwest MO cluster of warnings and LSRs. Both wind and hail forecasts also show increases in probabilities, though these increases are not as large as those seen for the tornado forecasts (Fig. 5.3g-r). While hail forecasts show the smallest increase in probabilities, there are probabilities present in areas where the model trained on LSRs-only did not produce probabilities. Despite still showing a weak signal, it is encouraging that WoFS-PHI is able to signal a weak hail threat as represented by warnings even though no hail was reported. Wind forecasts appear to have better spatial resolution with more of the higher-end probabilities associated with events when warnings are included in training. This is in part due to the changed event space. There are instances of warnings but where there were no LSRs, so these areas that appeared as false alarms when verifying on LSRs-only are now hits when verifying on LSRs and warnings. The version of WoFS-PHI that was trained on LSRs and warnings is better at highlighting these areas. So, while the warnings-trained version of WoFS-PHI appears to expand the area of highest probabilities, including wind-tagged warnings in the verification shows that what appears to be a forecast with low SRs for high-end forecasts if only LSRs are used to verify is actually a good representation of the severe hazard threat area.

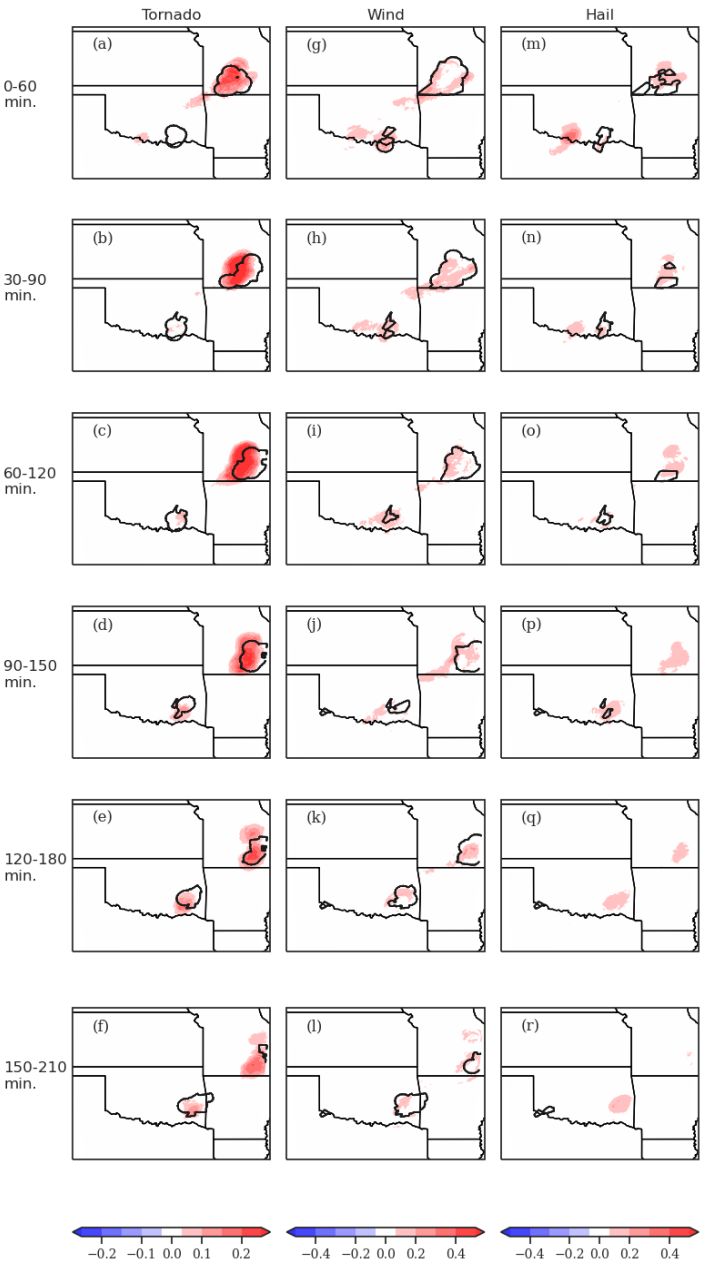

Figure 5.3: As in fig. 5.2, but between the LSR and warnings trained model (greater where red) and the LSR-only trained model (greater where blue). Both models were trained with both PS3 and TORP.

5.1.3 Impact of TORP without PS3

For the April 30, 2019 case, there is interestingly minimal change to the forecast probabilities when PS3 data is removed, even at early lead times (when PS3 is most important) for hail and especially wind (Fig. 5.4g-r). Since the WoFS and TORP model was very heavily reliant on WoFS, this suggests that this was a forecast driven by WoFS more than PS3. Given that this forecast was initialized at 0100z, these storms had been around for several hours prior to the 0100z WoFS initialization (this was confirmed by looking at earlier forecasts and LSRs/warnings on this date). It is known that the POD of storms for WoFS increases with storm lifetime (Guerra et al. 2022), so it makes sense that this forecast would be qualitatively good even in the absence of PS3 data.

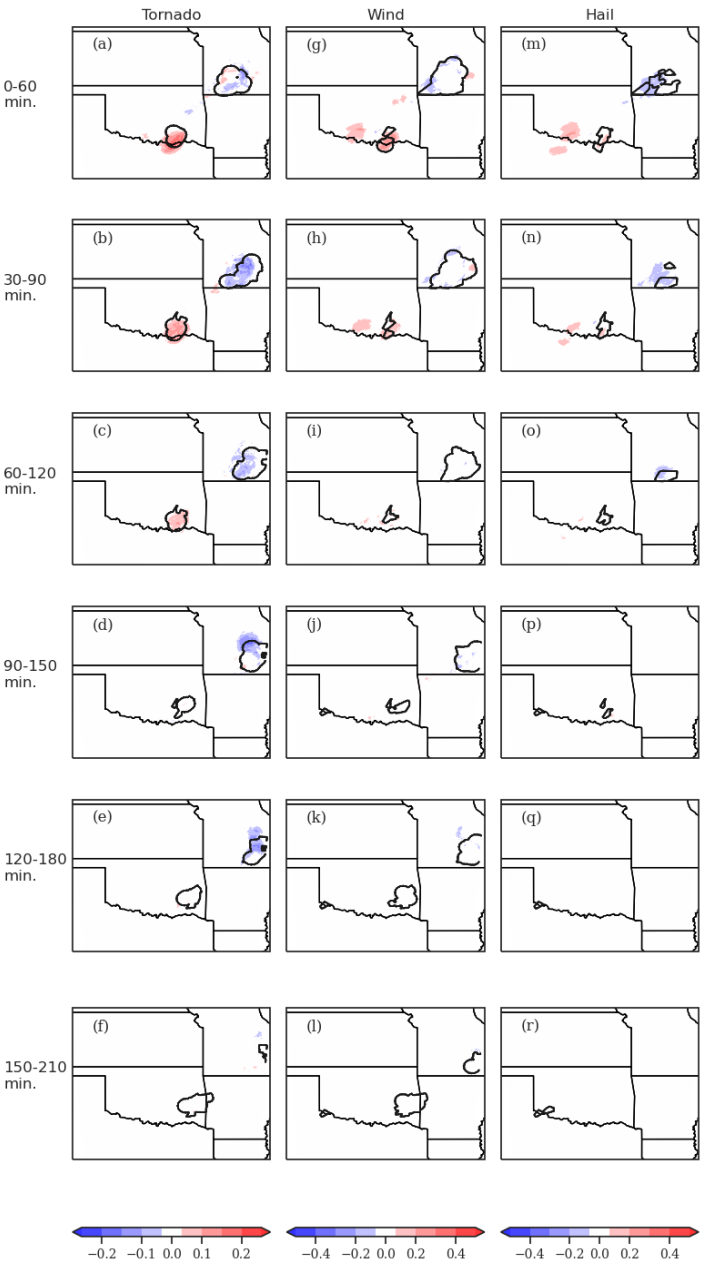

Figure 5.4: As in fig. 5.2, but the model trained on PS3 with TORP (greater where red) and the one trained on TORP without PS (greater where blue, both trained on LSRs and warnings).

The tornado forecasts show an interesting pattern between the isolated storm in far southern OK and the cluster in southwestern MO. The storm in far southern OK has higher probabilities when PS3 is included in training, but the cluster in MO has higher probabilities when only WoFS and TORP are used (Fig. 5.4a-c). From Fig. 2.3, it can be seen that there are no TORP objects associated with the southern OK storm, but there is a PS3 object. However, while the MO cluster has PS3 objects, the TORP objects have higher probability magnitudes (Fig. 2.3). The location and magnitudes of TORP objects combined with the knowledge that TORP is very influential (Fig. 4.28) when there are high-probability TORP objects available to use seems to indicate that TORP (rather than WoFS alone) is contributing to correctly raising the probabilities in southwestern MO above the level that they are at when PS3 is included in training. This is not to suggest that PS3 is not important or that TORP can replace PS3. It has already been made clear that PS3 is a vital contributor to WoFS-PHI and is substantially more useful than TORP. However, this is an example of the fact that while TORP is often unimportant due to unavailability, when there are $\geq 40\%$ TORP objects, they can provide useful information.

5.1.4 Baseline vs Final Model

A final comparison shows the overall change in probabilities from the baseline (PS2, no TORP, trained on LSRs-only) and the "final", best performing, model (PS3, with TORP, trained on LSRs and warnings). For this case, there is a clear increase in probabilities with very few areas of decreased probabilities (Fig. 5.5). Further, these increased probabilities occur predominantly in regions where there are LSRs and/or warnings, and the few regions of decreased probabilities occur in regions with no LSRs or warnings (Fig. 5.5). The improved skill associated with the PS3 and TORP

model trained on LSRs and warnings is clearly present in a qualitative analysis of this forecast, and the improvement is most apparent for tornado forecasts.

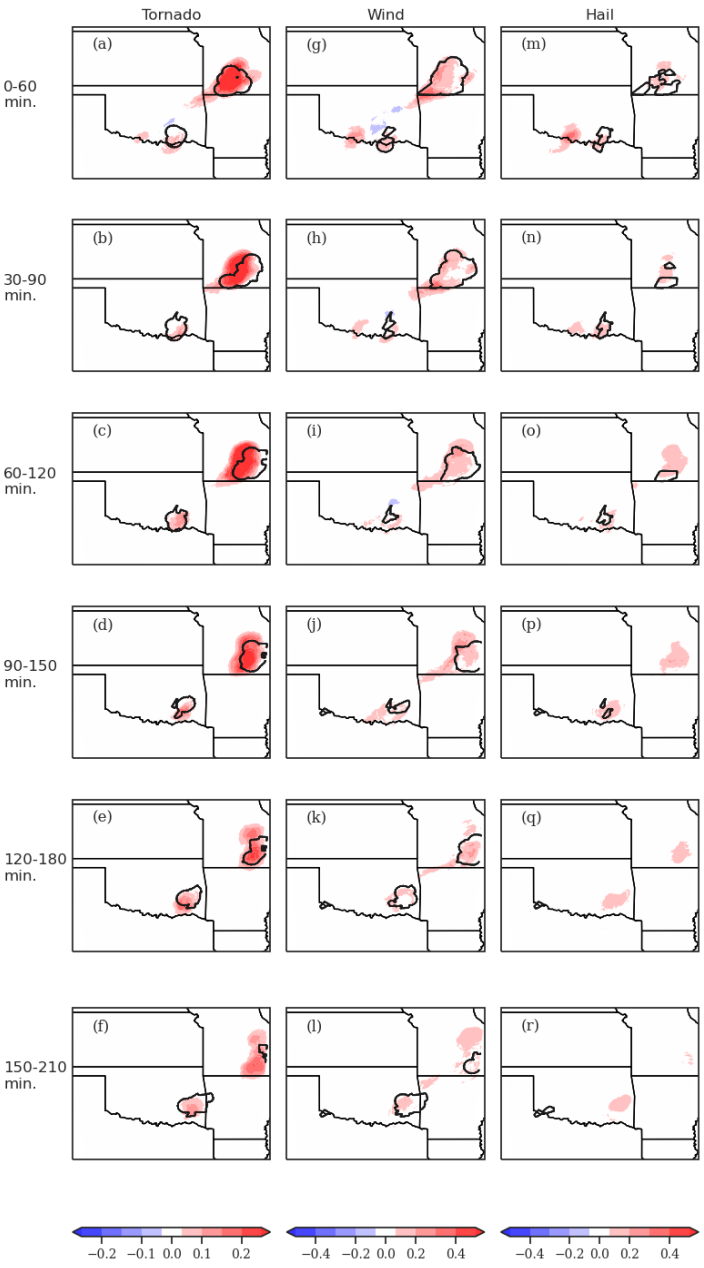

Figure 5.5: As in fig. 5.2, but between the model trained on PS3 with TORP and LSRs and warnings (greater where red) and the model trained on PS2 and no TORP and LSRs-only (greater where blue).

5.2 April 28, 2020

5.2.1 Impact of Changing the Predictor Dataset

For the April 28, 2020 forecast initialized at 2100z, there are no tornado LSRs in any of the forecast periods, yet there is noticeably more spatial coverage of probabilities than was seen with hail LSR forecasts on April 30, 2019 when there were no hail LSRs (Fig. 5.6a-f). Also interestingly is a bullseye of high tornado probabilities in southeast KS/northwest OK despite a lack of any LSR. The circular shape of the probablities with no probabilities surrounding it makes it appear as if a single PS2 object is highlighting a tornado threat, and WoFS-PHI is increasing the probabilities in turn. Also in this forecast, hail is highlighted as a bigger threat than wind at the early lead times with probabilities focused over southeast KS and northeast OK, but wind has higher probabilities at later lead times (Fig. 5.6g-r). A similar pattern to wind probabilities from the April 30, 2019 forecast appears for both wind and hail forecasts here with decreasing spatial resolution with lead time. While in the April 30, 2019 forecast, wind probabilities did not decrease (much) with lead time, the maximum wind probabilities appear to actually increase with lead time in this forecast, and certainly the spatial extent of wind probabilities increases with lead time with a line of wind probabilities extending from central OK to southwest MO (Fig. 5.6g-l). This is in contrast with hail probabilities that decrease in magnitude with lead time (Fig. 5.6g-l). The patterns shown for both wind and hail are consistent with the performance metrics of hail forecast quality degrading faster with lead time than wind (Fig. 3.3a,b).

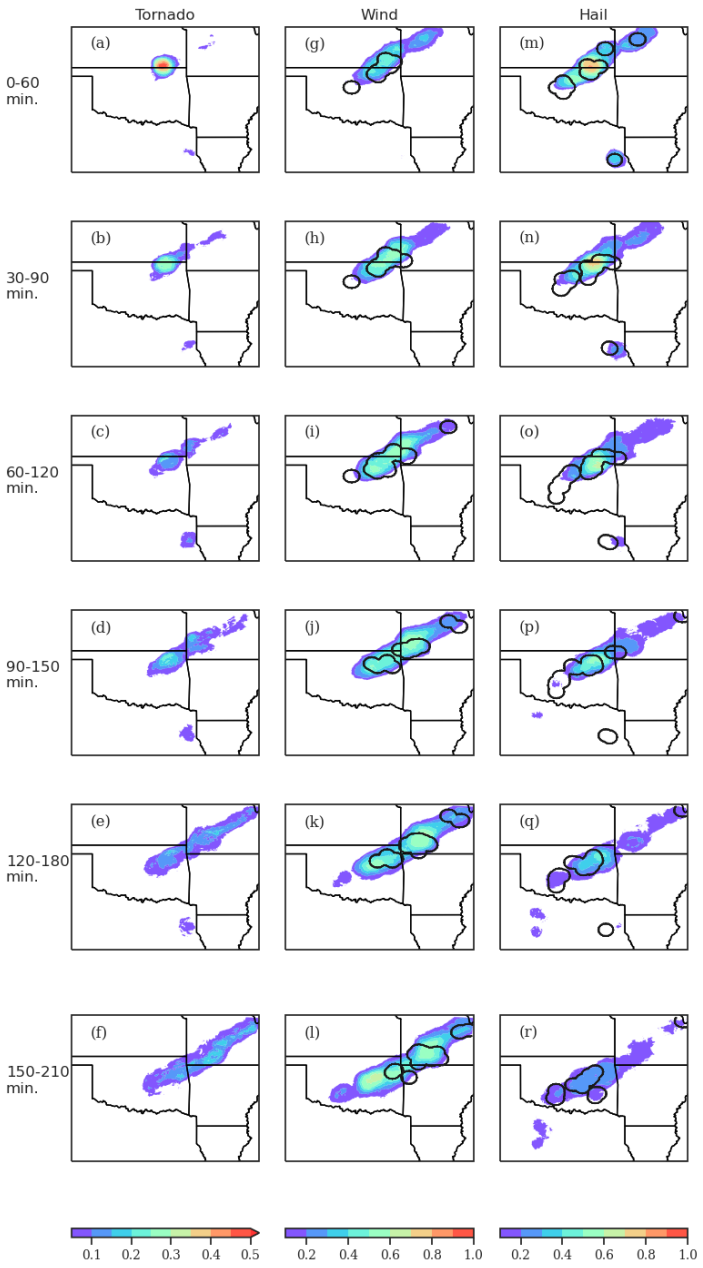

Figure 5.6: As in fig. 5.1, but for April 28, 2020 from 4:30pm to 8pm local time.

The forecasts for April 28, 2020 appear visually very similar for wind when PS3 and TORP are included in training (Fig. 5.7g-l). Hail forecasts also appear very similar with a slightly higher peak probability magnitude in the 0-60 minute window (Fig. 5.7m-r). A notable difference in the first two tornado lead times shows that the tornado probability hotspot that existed with PS2 no longer exists when PS3 and TORP data are included (Fig. 5.7a,b). This lends further credence to the idea that this lone hotspot was perhaps due to a PS2 object incorrectly signaling high tornado likelihood where the same object generated by PS3 does not show as high of probabilities.

There is also a notable line of hail LSRs that extents from central OK southward at 90 minutes of lead time (Fig. 5.7o). Even with the new observational data, this line of reports is not highlighted by WoFS-PHI at all. However, there is a slight signal for increased hail probabilities in this region at the latest lead times with the new observation-based data (Fig. 5.7q,r). It seems likely that this area represents a line of storms that had not yet initiated by 2100z (thus prohibiting any benefit from PS3 or TORP), and WoFS likely did not resolve the initiation of these storms well, or perhaps mistimed the initiation given that the region is highlighted with low hail probabilities at longer lead times after the LSRs appear.

Additionally, it is clear that the increased hail probabilities associated with the new observational data increase probabilities almost exclusively in areas within 39 km of an LSR despite the gaps between the LSRs in the threat area. While the overall hail probability swath covers the whole area, the localized probability increases with the new observational data contributed to correctly increasing hail probabilities in targeted areas.

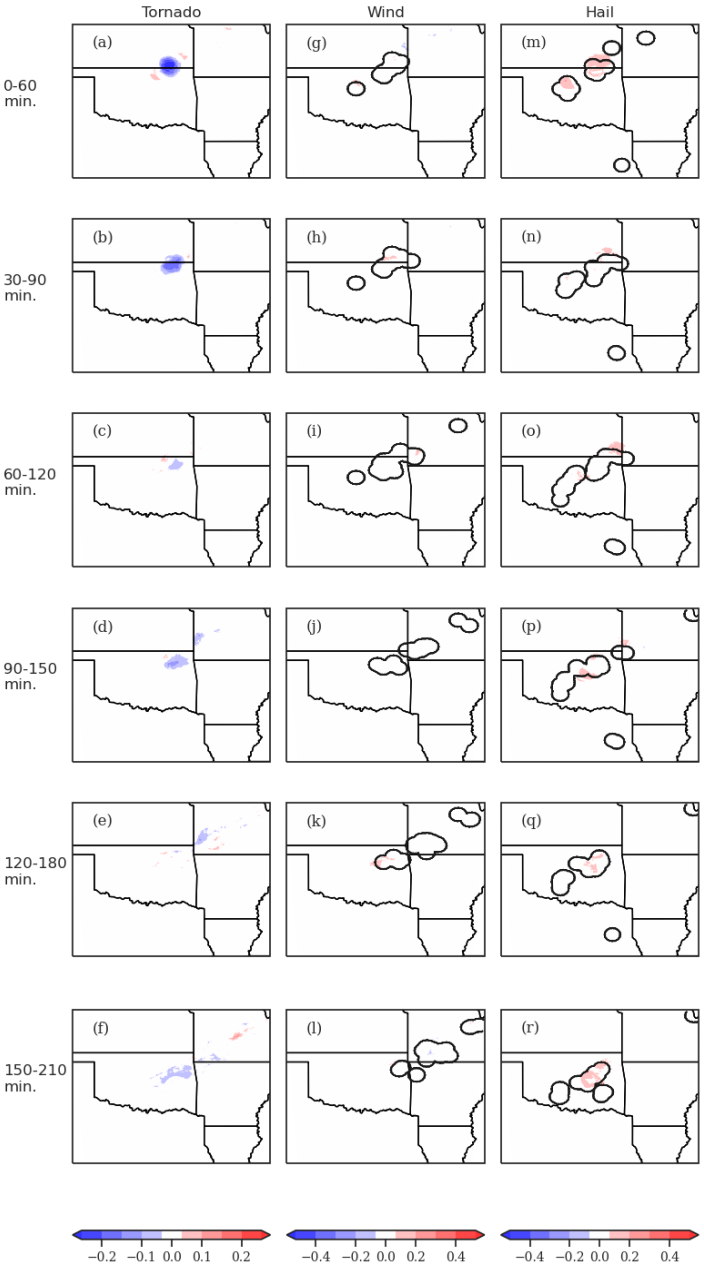

Figure 5.7: As in fig. 5.2, but for April 28, 2020 from 4:30pm to 8pm local time.

5.2.2 Impact of Training on Warnings

An expected pattern appears in the April 28, 2020 case. All three hazards see substantial increases in probabilities at all lead times, and wind and hail see the largest increases (Fig. 5.8). For tornado forecasts, there were no LSRs, but there were several warnings and at least one warning at each lead time (Fig. 5.8a-f). Also, the small region southeastern KS/northeastern OK that had high tornado probabilities from the model trained using PS2 but low probabilities from the model trained with PS3 turns out to have multiple warnings in the vicinity during the 0-60 and 30-90 minute periods (Fig. 5.8a,b). This is an example of a time when verifying on LSRs can be misleading. It would appear that the model trained on PS2 is worse in this instance when verifying on LSRs-only since it produces a region of high probability with no LSR to match it while the model trained with PS3 produces much lower probabilities (Fig. 5.7). However, when warnings are included in training and verification, the probabilities are correctly raised relative to the LSR-only trained model to reflect the tornado threat even if no tornado was reported (Fig. 5.8a,b).

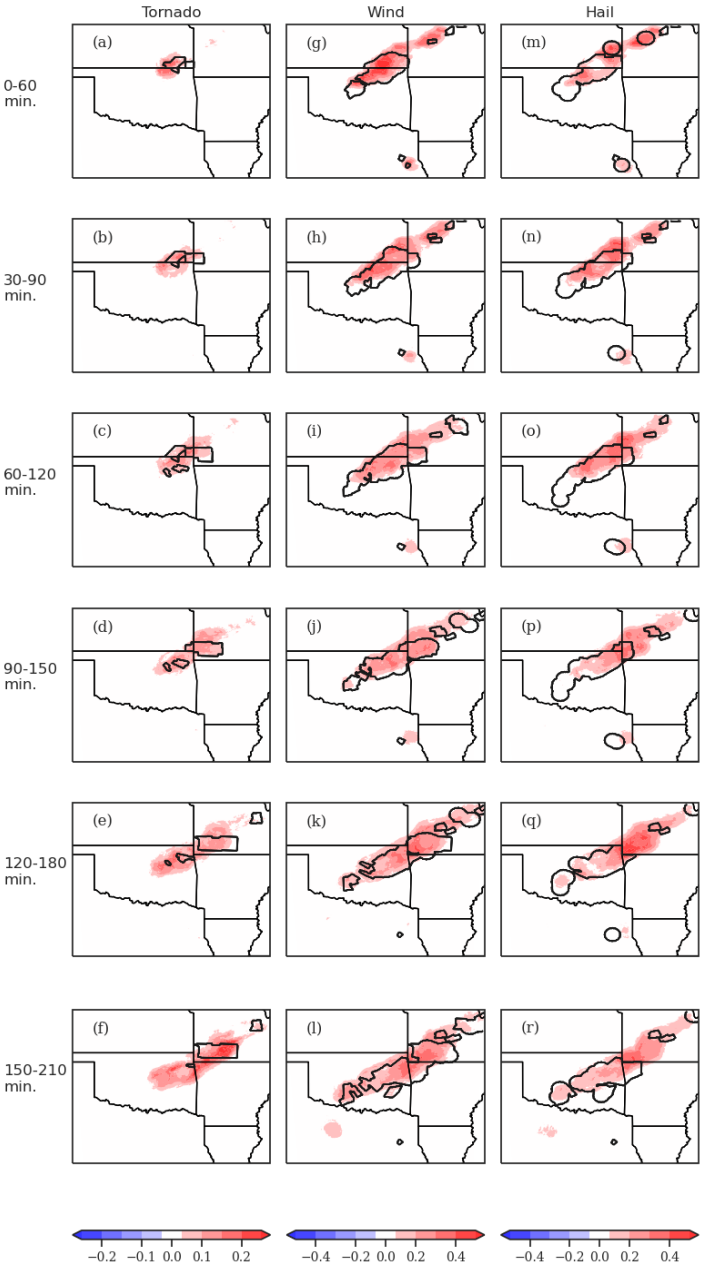

Figure 5.8: As in fig. 5.3, but for April 28, 2020 from 4:30pm to 8pm local time.

Wind forecasts show impressive resolution at all lead times for the April 28, 2020 forecast. The combination of large increases in probabilities with the increase in spatial coverage in high probabilities makes it especially impressive that nearly all of the increased probabilities fall within the bounds of the LSR or warning area. Not only this, but almost the entire area of LSRs and warnings sees increased wind probabilities with the warnings-trained model (Fig. 5.8g-l).

The hail threat (as determined by LSRs and warnings) appears to be mostly contained within central and eastern OK (and southeastern KS at early lead times), and while there are some warnings southwestern MO, they are not nearly as expansive as those over OK (Fig. 5.8m-r). However, similarly elevated hail probabilities to those seen over OK can be found over MO (Fig. 5.8m-r). The impact of training on warnings also appears to raise hail probabilities slightly more at later lead times, although the sharpness increases are more focused in areas with LSRs and/or warnings at early lead times (Fig. 5.8m-r).

The overall impact of LSRs and warnings on training in these forecasts supports the findings that overall probabilities are greatly increased (especially the April 28, 2020 case), but the relatively small impact from including warnings in training for hail and wind forecasts in the April 30, 2019 case is a reminder that the impact of training on warnings is not uniform across all forecasts. An investigation into the potential drivers of when LSRs are a better or worse representation of the severe threat (or if this is simply random variability) is left to future work. Identifying such a pattern (if one exists) would be helpful for identifying potential biases in warnings and for identifying scenarios in which training against LSRs-only may be more or less acceptable.

5.2.3 Impact of TORP without PS3

There are generally higher probabilities from the model that includes PS3 in training for the April 28, 2020 case, but there is an especially notable difference in the far northeastern part of the domain in western and central MO in the first two forecast periods. This suggests that PS3 was especially useful in these regions (Fig. 5.9g,h,m,n). This appears to be a case where where WoFS is handling the southern cluster of storms well but not resolving the nothern cluster as well, despite the northern cluster not being an example of new convection. This forecast is a good example of how PS3 is still important to forecasts even after storms have persisted for several hours. The strengths of WoFS and PS3 that have been highlighted in different parts of the example forecasts make a strong case that both are valuable and needed to produce the best WoFS-PHI forecasts.

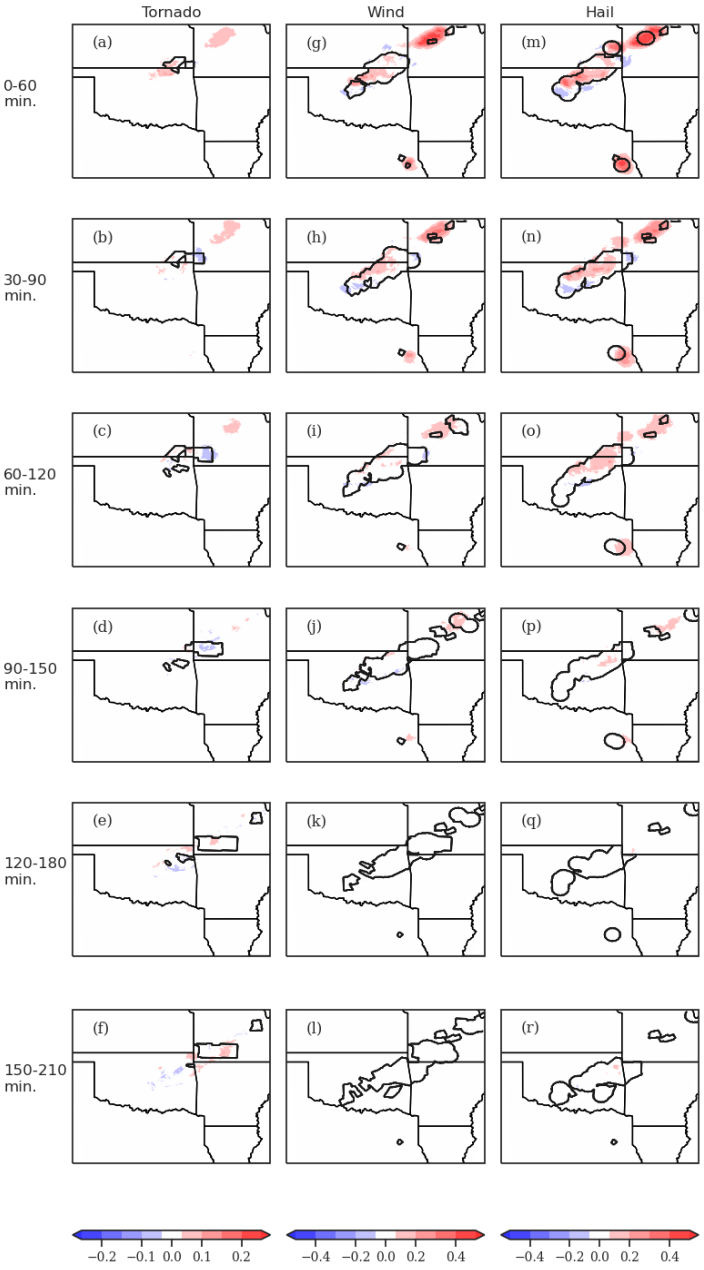

Figure 5.9: As in fig. 5.4, but for April 28, 2020 from 4:30pm to 8pm local time.

Tornado forecasts for this case are least impacted by the loss of PS3 (Fig. 5.9a-f). This is perhaps surprising given how important PS3 was to tornado forecasts especially, but a closer look reveals that neither PS3 nor TORP contributed much in this case. While there were PS3 objects in the vicinity of the warnings in the 0-60 minute period, they all had very low tornado probabilities (Fig. 5.10). Similarly, most TORP objects were low probability objects as well (Fig. 5.10c). Low probability TORP objects were shown to have nearly no influence on RF tornado probabilities, even with the WoFS and TORP without PS3 model (Fig. 4.28). Thus, it seems that for this forecast, regardless of whether PS3 predictors (or TORP predictors) were available, this was going to be a mainly WoFS-driven forecast.

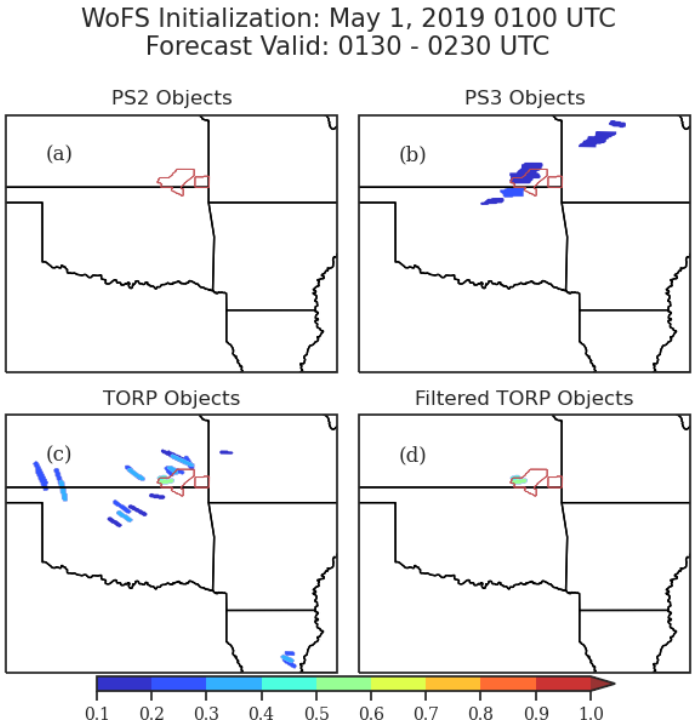

Figure 5.10: As in fig. 2.3 showing the grid-mapped forecast objects from PS2 (a), PS3 (b), TORP (c), and 40%+ probability TORP objects (d). These objects are from the forecast using a WoFS initialization from 4:00pm local time on April 28, 2020. The mapped objects represent their swaths for the 4:30pm to 5:30pm local time forecast.

5.2.4 Baseline vs Final Model

For the April 28, 2020 case, the comparison between the baseline (PS2, no TORP, trained on LSRs-only) and the "final", best performing, model (PS3, with TORP, trained on LSRs and warnings) shows similar results to the April 30, 2019 case. There is again a clear increase in probabilities focused in regions with LSRs and/or warnings with very few areas of decreased probabilities (Fig. 5.11). While the April 30, 2019 case showed the most improvement for tornadoes, the April 28, 2020 case shows the most improvement for hail and wind forecasts. That both forecasts show substantial improvement and for different hazards is encouraging as the dependence of WoFS and PS3/TORP data seems to vary across the domain and between forecasts, yet regardless of lead time or local predictor importance, the aggregated effect of the changes discussed in this paper lead to improved spatial forecasts of severe weather (i.e., increased probabilities where warnings or LSRs occurred and decreased probabilities where warnings or LSRs did not occur). Further, not only is there vastly increased sharpness in the final model, these higher probabilities are almost exclusively within or at least near LSRs and warnings. While a look at individual forecasts provides only a qualitative and subjective assessment, these and many other forecasts from the dataset provide clear supporting evidence of much of the behavior seen in the more abstract verification and explainability discussion. These forecasts show visual representations of the final model's increased sharpness, POD, SR, and reliability.

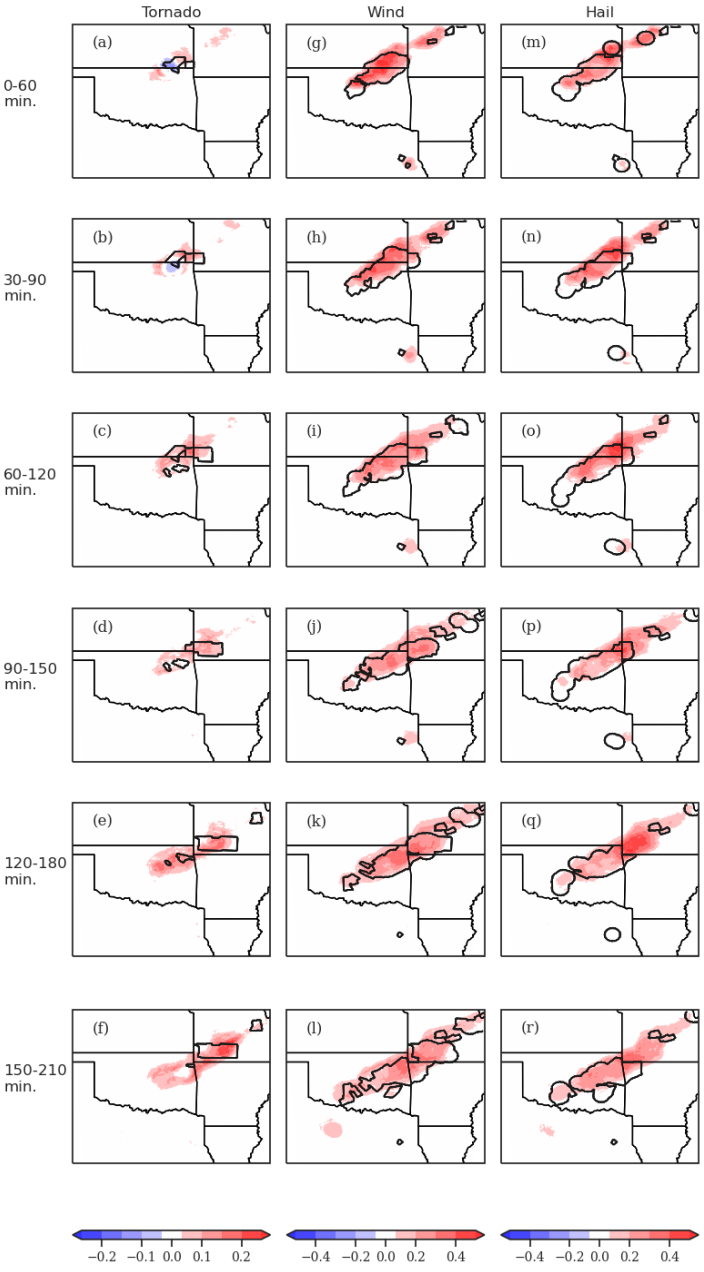

Figure 5.11: As in fig. 5.2, but for April 28, 2020 from 4:30pm to 8pm local time.

Chapter 6

Discussion and Conclusion

Various experiments were conducted to assess methods of improving WoFS-PHI forecasts as detailed in Table 2.3. First, changes to the predictor dataset were made to determine the impact of including TORP data, using PS3 instead of PS2, and the impact of using both PS3 and TORP data compared to just PS2 (with all models including WoFS predictors). Next, the impact of changing the target dataset was also assessed through changing the target dataset to include warnings. The final experiment was to train a model with WoFS and all TORP predictors but no PS3 to find out how much of a substitute TORP could be for PS3 data.

These experiments were conducted through verification comparisons between models using BSS and performance and attributes diagrams. The behavior of individual predictors and predictor fields was also assessed through RTI (for both sets of experiments) and ALE analyses (for the experiments testing changes to the predictor datasets. Lastly, two example forecasts were used as case studies to identify how the verification and explainability results manifest in real forecasts.

6.1 Q1: What is the Impact from Changing the Predictor Dataset?

6.1.1 Q1a: What is the impact of adding TORP predictors to training?

When comparing the baseline (PS2 without TORP) to the PS2 with TORP model, the biggest impact was seen for hail forecasts (Fig. 3.1a) despite TORP being a *tornado* detection algorithm. However, while hail forecasts saw the largest increase in forecast quality from including TORP in the predictor dataset, the magnitude of impact was relatively similar across all hazards, and the confidence intervals overlapped at all lead times for all hazards (Fig. 3.1).

RTI analyses provide some explanation as to why there was minimal increase in forecast quality with the addition of TORP predictors. TORP predictors proved to have minimal importance compared to WoFS and ProbSevere predictors for all hazards (Figs. 4.4, 4.5, 4.6). However, ALE analyses show that TORP predictors can provide substantial influence when available through the maximum reflectivity associated with each TORP object for hail and wind forecasts (Figs. 4.13, 4.16) and TORP probability for tornado forecasts (Fig. 4.19). This suggests that there are simply very few TORP objects available. While most WoFS-PHI forecasts are unable to use TORP predictors due to lack of availability (thus leading to low importance in RTI), TORP predictors are quite useful when they are available.

TORP predictors also are only substantially influential for maximum reflectivity $\geq$ 50 dBZ for hail and wind forecasts (Figs. 4.13), and TORP probability $\geq$ 40% for tornado forecasts (Fig. 4.19). Since only higher-end TORP objects are influential according to ALE analyses, this would both support the hypothesis that the lack of

availability of TORP objects is driving its lack of importance and contribution to forecast quality, but it also helps to explain why only higher-end forecasts (of all hazards) see increased forecast quality. In chapter 2, it was discussed that the method for ALE analysis for TORP objects was adjusted since the distribution of TORP objects was too heavily skewed towards "0" values for a curve to even be generated. This is further evidence of the lack of TORP predictor availability.

Overall, TORP has a very small (though positive) impact on WoFS-PHI forecast quality. This impact is limited by the lack of TORP objects available to WoFS-PHI as it appears to be impactful when it is available (Figs. 4.13, 4.16, 4.19). A noted weakness in the training dataset was the lack of days with large numbers of tornadoes. Including more tornadoes in the training dataset could help to increase the number of TORP objects and increase it's overall importance to forecasts. For hail and wind forecasts, use of low level radar data across the entire domain could prove to be a skillful set of predictors given the influence of TORP reflectivity when available. Either regridded radar data across the domain or a similar algorithm to TORP that detects hail and/or wind could help to reduce the constraint of the lack of data availability imposed by TORP.

6.1.2 Q1b: What is the impact of using PS3 predictors instead of PS2?

When comparing the PS2 with TORP model to the PS3 with TORP model, there are again minimal differences in forecast quality. The biggest increase in skill is for tornado forecasts followed by hail and wind (Fig. 3.2). For all hazards (including tornadoes, which see the biggest impact from using PS3 instead of PS2), there is

minimal change to the overall importance of ProbSevere hazard probabilities between PS2 and PS3 in RTI analyses (Figs. 4.1, 4.2, 4.3, 4.4, 4.5, 4.6).

However, ALE analysis again provides more insight as to what is driving the marginal increases in forecast quality. For hail forecasts, the influence of hail ProbSevere probabilities increases substantially when using PS3 instead of PS2, but the influence of wind ProbSevere probabilities decreases (Fig. 4.13). The influence of tornado ProbSevere probabilities increases, but it is still quite low and its influence increases much less than hail ProbSevere probabilities. Wind forecasts are influenced approximately equally by PS2 and PS3 wind probabilities, but hail and tornado ProbSevere probability influences decrease when using PS3 instead of PS2 (Fig. 4.16). Similarly, tornado PS3 probabilities are more influential than PS2, and hail and wind PS3 probabilities are less influential than PS2 (Fig. 4.19). These patterns indicate that PS3 provides better discrimination between hazards than PS2, and the fact that this pattern is strongest for tornadoes helps to explain why tornadoes see the biggest increase in forecast quality from PS3.

6.1.3 Q1c: What is the overall impact of using TORP and PS3

The aggregate influence of TORP and PS3 together provides more significant increases in skill for hail and tornadoes in the $0-60$ minute period (Fig. 3.3). This increased forecast quality appears to be approximately the sum of the TORP and PS3 effects individually, which makes sense. However, for both TORP and PS3 (and the combined impact), this impact is restricted to earlier lead times with impact dropping off quickly with lead time, especially for TORP. This makes sense given prior work (Loken et al. 2022b, in review) and logic given that PS3 is meant to

forecast severe potential within the next hour, and TORP provides instantaneous tornado potential. Neither are equipped to forecast at long lead times, and this is reflected in the neglible forecast quality increases at longer lead times and the increased importance (Figs. 4.4, 4.5, 4.6) and influence (Figs. 4.14, 4.15, 4.17, 4.18, 4.20, 4.21) of WoFS predictors at later lead times.

6.1.4 Q1d: What is the impact of TORP when neither PS2 nor PS3 are used?

Across all three hazards, the WoFS and TORP (without ProbSevere) model is heavily reliant upon WoFS rather than TORP data at all lead times (Figs. 4.10, 4.11, 4.12). The total contribution by WoFS predictors is approximately 95% for 0-60 minute forecasts (slightly less for wind) and increases with lead time to near 99% for 150-210 minute forecasts. It appears that TORP does very little to replace the observational data void left by the removal of PS3 data, and this is reflected in the large drop in skill at early lead times when PS3 data is removed (Fig. 3.15). Attributes diagrams suggest that this drop in performance at early lead times is driven by a decrease in resolution rather than reliability (Figs. 3.16, 3.18, 3.20), and this is consistent with the performance diagrams showing a predominantly downward (as opposed to leftward) shift which indicates a decrease in the POD for a given SR (Figs. 3.17, 3.19, 3.21). This direction in which the performance diagram curve shifts is important because it suggest that the main weakness of the TORP-only model is in missed events that lower its POD rather than false alarms that lower its SR.

Despite TORP's lack of overall importance, it has a substantial contribution to hits (considering the lack of overall importance) for all hazards at early lead times (Figs. 4.10, 4.11, 4.12). This suggests that when TORP is available, it helps to

increase confidence in higher-end forecasts. This claim is supported by ALE analysis that shows TORP reflectivity to have an extremely high ALE for reflectivity values of 50+ dBZ for hail and wind forecasts (Figs. 4.22, 4.25). Similarly, TORP probability has a nearly equal ALE value to when ProbSevere is excluded to what PS3 tornado probability had when ProbSevere was included (Fig. 4.28). All of this suggests that the rarity of TORP objects is the driving contributor to why TORP data is not important to forecasts in general. The missed events that lower the skill of the WoFS and TORP model are likely a result of the lack of TORP objects that ProbSevere highlighted when it was available.

There are multiple potential reasons for this lack of TORP objects. One is that the nature of TORP is to be a diagnostic algorithm rather than a prognostic algorithm. Thus, where ProbSevere might highlight a non-tornadic storm to have tornadic potential within the next hour, TORP may not highlight that same storm since it is not immediately tornadic. Second, a simple analysis of the percentage of grid points covered by TORP objects revealed that while the total number of TORP objects did not lag PS3 substantially, the distribution of TORP objects is heavily skewed towards the lowest probabilities such that most TORP objects are not influential to forecasts. However, the influence on hail and wind forecasts, especially, suggests that including radar data across the domain (as opposed to only where there is an immediate tornadic threat) could substantially improve forecast performance. For tornado forecasts, this may still help as radar data would not be limited to regions of immediate tornadic potential. Theoretically, WoFS-PHI could learn patterns to allow it to raise probabilities in areas of tornadic potential associated with radar predictors. TORP radar predictors are also associated with the 0.5° scans, and using higher-level scans may help identify hail cores aloft or mesocyclones in storms that have not developed to lower levels yet.

Lastly, a more robust analysis of the influence of TORP without PS3 would also train a model using only WoFS predictors as a comparison. This is done in Loken et al. (2022b, in review) but only for 30-minute durations. Due to time and computational constraints, this experiment was not conducted with this work, but it would provide insight as to the skill benefit provided by TORP when PS3 is not used even considering the limitations of TORP's lack of availability in many cases.

6.2 Q2: What is the impact of using a target dataset that includes hazard-specific warnings in addition to LSRs?

WoFS, PS3, and TORP were all used to train models that were verified on LSRs-only and LSRs and warnings. The goal of including warnings in the target dataset during training was to help mitigate various spatiotemporal biases assoicated with LSRs (Trapp et al. 2006; Allen and Tippet 2015; Smith et al. 2013; Edwards et al. 2018). While it remains unclear whether warnings represent higher quality "truth" data, training on warnings did reliably increase the frequency of high-end probabilities in forecasts for all hazards (Figs. 3.9, 3.11, 3.13). In turn, this significantly increases BSS at all lead times for hail and wind forecasts (Fig. 3.8a,b) and most lead times for tornado forecasts (Fig. 3.8c). It also increases the SR and especially probability of detection (POD) at each forecast threshold (Figs. 3.10, 3.12, 3.14). Further, both reliability and resolution are improved across all hazards, even in cases where there was under forecasting for the LSR-only trained models (in contrast to the effect of the changed/added predictor datasets that slightly worsened instances of under forecasting).

As noted in chapter 3, the impact of event climatology must be considered when comparing performance metrics, especially for BSS. The fact that forecast quality improvement is relatively consistent with lead time for all hazards also suggests that severe/tornado warnings are *more predictable* than severe/tornado LSRs. Further work should explore the reason for why the inclusion of warnings in the target dataset leads to improved forecasts. It is hypothesized that warnings are truly a more predictable target and have fewer biases than LSRs, but it is also possible that the higher climatology of LSRs and warnings compared to LSRs-only is the main driver behind the improved forecast performance. To test this, the LSR and warnings dataset could be sampled such that its event climatology matches that of the LSR-only dataset. Then, this sampled dataset could be used to train a new model to compare to the LSR-only model. These models could be verified against the LSR-only dataset, the reduced LSR and warnings dataset, and the full LSR and warnings dataset to see how the two models compare in predicting each type of target. The degree to which the difference in BSS is reduced would represent how much of the BSS improvement displayed here is due to the simply increased event climatology associated with including warnings in the training dataset. Any BSS improvement that remains would be indicative of including warnings leading to a higher quality "truth" dataset.

For now, the hypothesis of warnings being a more predicable target is supported by the explainability of the LSR-only and LSR and warning models. Across all hazards, the fractional contribution by all predictors to hits increases when the model is trained on warnings, but the fractional contribution to misses increases only marginally (Figs. 4.4, 4.5, 4.6, 4.7, 4.8, 4.9). Even though more events allows for the potential for both more hits and more misses, the result is dominated by an increase in the number of hits. Both WoFS and PS3 (and to a slight extent, TORP)

predictors show similar increases in contributions to hits, but not nearly as much misses. This result shows that a warning-or-LSR event is better correlated with the predictors in WoFS-PHI than LSRs alone and perhaps a better representation of severe weather. Further, among the large changes to predictor usage is a decrease in the usage of geographical predictors when WoFS-PHI was trained on LSRs and warnings (especially for wind and hail forecasts; Fig. 4.4, 4.5, 4.7, 4.8). This provides further evidence that at least one of the known biases in LSRs (Smith et al. 2013; Allen and Tippet 2015; Edwards et al. 2018) is being addressed by including warnings in training events.

However, as mentioned, warnings are not perfect, and biases associated with warnings can be identified in these results as well and should be acknowledged. Most notable among these is the coupling of wind and hail in forecasts. The most obvious representation of this is in the usage of PS3 wind probabilities for hail forecasting and in PS3 hail probabilities for wind forecasting. PS3 hail probabilities are used more for wind prediction than PS3 wind probabilities are used for hail prediction. Further, the contribution to hits (relative to the increased contribution to hits from all predictors) of PS3 hail probabilities for wind prediction increases substantially when warnings are included in the target dataset (Figs. 4.4, 4.5, 4.7, 4.8).

It has been shown that PS3 wind probabilities have been shown to be less useful to early lead time wind forecasts than PS3 hail probabilities are for hail forecasts and that WoFS is more important for early-lead-time wind forecasts than hail or tornado forecasts. It is possible then that a coupling of hail and wind predictors for hail forecasting would manifest more in WoFS predictors than PS3. There is some indication of this as the importance of flash extent density (an indication of general severe threat rather than a hazard-specific predictor) increases as does the influence of 80 m wind (a wind-specific predictor; Figs. 4.4, 4.7). However, even among WoFS

predictors known to be associated with wind forecasts, the increased influence on hail forecasts is not substantial. It is also possible that warnings with severe hail tags are less likely to have severe wind tags than warnings with severe wind tags are to have severe hail tags. Regardless, the bias created by the use of combined severe hail and wind warnings is notable in the RTI and ALE analyses of WoFS-PHI hail and especially wind forecasts.

Another bias that must be acknowledged is the use of severe thunderstorm warnings with "tornado possible" tags to represent tornadoes. This bias appears to be present in the explainability of tornado forecasts. The RTI analysis show an increased contribution from PS3 wind probabilities when WoFS-PHI is trained to predict warnings in addition to LSRs (4.6, 4.9). A version of WoFS-PHI trained only on tornado LSRs and warnings without "tornado possible" warnings was not evaluated in detail. It is thus an open question whether this change in use of PS3 predictors is entirely due to the use of severe thunderstorm warnings in the training data.

6.3 Conclusions

While the changes to the predictor dataset provided benefits to WoFS-PHI skill (especially for hail and tornado forecasts and at early lead times), the biggest impact was seen when warnings were included in the target dataset with forecasts being substantially sharper and having higher BSS, POD, and SR values at all lead times. While further work is needed to address the most important drivers of this improvement (i.e. the impact of an increased event climatology compared to the differences in the distribution of warnings verus LSRs), it is clear that the use of the LSR and warnings trained model for real-time implementation is justified.

While TORP and PS3 are not available for operational usage in WoFS-PHI at the time of writing, these findings have led to the currently operational version of WoFS-PHI being trained on LSRs and warnings as described here. Additionally, a 2-hour forecast duration product has been implemented operationally (in addition to the 1-hour product described here). The version of WoFS-PHI that uses TORP and PS3 predictors is expected to be deployed when these tools become available to WoFS-PHI in real time. This project shows the value of continuing to develop operational forecast tools both for improved forecast performance and for displaying new techniques to benefit other products.

Reference List

Allen, J. T., and M. K. Tippet, 2015: The characteristics of united states hail reports: 1955–2014. *Electronic J. Severe Storms Meteor.*, **10**, 1–31, https://doi.org/10.55599/ejssm.v10i3.60.

Auer Jr., H., A., 1994: Hail recognition through the combined use of radar reflectivity and cloud-top temperatures. *Mon. Wea. Rev.*, **122**, 2218–2221, https://doi.org/10.1175/1520-0493(1994)122⟨2218:HRTTCU⟩2.0.CO;2.

Bauer, P., 2024: What if? numerical weather prediction at the crossroads. *J. Eur. Meteor. Soc.*, **1**, https://doi.org/10.1016/j.jemets.2024.100002.

Benjamin, S. G., and Coauthors, 2016: A north american hourly assimilation and model forecast cycle: The rapid refresh. *Mon. Wea. Rev.*, **144**, 1669–1694.

Berry, K., and Coauthors, 2023: 2023 experimental warning program experiment summary. Tech. rep., NOAA Experimental Forecast Program of the Hazardous Weather Testbed.

Blair, S. F., D. R. Deroche, J. M. Boustead, J. W. Leighton, B. L. Barjenbruch, and W. P. Gargan, 2011: A radar-based assessment of the detectability of giant hail. *Electron. J. Severe Storms Meteor.*, **6**, 1–30, https://doi.org/10.55599/ejssm.v6i7.34.

Blair, S. F., and Coauthors, 2017: High-resolution hail observations: implications for nws warning operations. *Wea. Forecasting*, **32**, 1101–1119, https://doi.org/10.1175/WAF-D-16-0203.1.

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

Brooks, H. E., and Coauthors, 2019: A century of progress in severe convective storm research and forecasting. *Meteor. Monogr.*, **59**, 18.1–18.41, https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0026.1.

Cintineo, J. L., M. J. Pavlonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: Noaa probsevere v2.0—probhail, probwind, and probtor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Cintineo, J. L., M. J. Pavolonis, and J. M. Sieglaff, 2024: Probsevere version 3: Improved exploitation of data fusion and machine learning for nowcasting severe weather. *Wea. Forecasting*, **39**, 1937–1958, https://doi.org/10.1175/WAF-D-24-0076.1.

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

Clark, A. J., and Coauthors, 2023: The first hybrid noaa hazardous weather testbed spring forecasting experiment for advancing severe weather prediction. *Bull. Amer. Meteor. Soc.*, **104**, E2305–E2307, https://doi.org/10.1175/BAMS-D-23-0275.1.

Clark, A. J., and Coauthors, 2024: Spring forecasting experiment 2024 preliminary findings and results. Tech. rep., NOAA Experimental Forecast Program of the Hazardous Weather Testbed.

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

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. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Flora, M. L., P. Skinner, C. K. Potvin, B. Matilla, and A. Reinhart, 2025: Assessing the impact of biased target variables on machine learning models of severe hail. *Wea. Forecasting*, https://doi.org/10.1175/WAF-D-24-0051.1.

Gesell, I. W., 2020: Verification of the tornado and lightning plumes and evaluation of a new kernel for the tornado and lightning plumes. M.S. thesis, University of Oklahoma.

Guerra, J. E., P. S. Skinner, A. Clark, M. Flora, B. Matilla, K. Knopfmeier, and A. E. Reinhart, 2022: Quantification of nssl warn-on-forecast system accuracy by storm age using object-based verification. *Wea. Forecasting*, **37**, 1973–1983, https://doi.org/10.1175/WAF-D-22-0043.1.

Harrison, D. R., A. McGovern, C. D. Karstens, I. L. Jirak, and P. Marsh, 2022: Evaluation of first-guess watch guidance in the 2022 HWT spring forecast experiment. *30th Conf. on Severe Local Storms*, Santa Fe, NM, Amer. Meteor. Soc., http://dharrisonwx.com/papers/2022_SLS_Extended_Abstract.pdf.

Hazardous Weather Testbed, 2021: HWT History. NOAA, accessed 2025, https://hwt.nssl.noaa.gov/history/.

Heinselman, P. L., and Coauthors, 2023: Warn-on-forecast system: From vision to reality. *Wea. Forecasting*, **39**, 75–95, https://doi.org/10.1175/WAF-D-23-0147.1.

Hempel, B., 2022: Nadocast—conus severe weather probabilities via feature engineering and gradient boosted decision trees. Github, accessed 2024, https://github.com/brianhempel/nadocast.

Hill, A. J., R. C. Schumacher, and I. L. Jirak, 2023a: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Wea. Forecasting*, **38**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Hill, A. J., R. S. Schumacher, and I. J. Jirak, 2023b: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Wea. Forecasting*, **38**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Ho, T. K., 1998: The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**, 832–844, https://doi.org/10.1109/34.709601.

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. *Wea. Clim. Soc.*, **3**, 128–140, https://doi.org/10.1175/2011WCAS1076.1.

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

Iowa State Univeristy, 2021: Iowa environmental mesonet, 2021: Archived nws watch/warnings. Iowa Environmental Mesonet, accessed 2021, https://mesonet.agron.iastate.edu/request/gis/watchwarn.phtml.

Jirak, I. L., M. S. Elliot, R. S. Karstens, C. D. ad Schneider, P. T. Marsh, and W. F. Bunting, 2020: Generating probabilistic severe timing information from spc outlooks using the href. *100th AMS Annual Meeting*, Boston, MA, Amer. Meteor. Soc., https://ams.confex.com/ams/2020Annual/webprogram/Paper367695.html.

Johnson, A. W., and K. E. Sugden, 2014: Evaluation of sounding-derived thermodynamic and wind-related parameters associated with large hail events. *Electronic J. Severe Storms Meteor.*, **9**, 1–42, https://doi.org/10.55599/ejssm.v9i5.57.

Jurkovic, P. M., N. S. Mahovic, and D. Pocakal, 2015: Lightning, overshooting top and hail characteristics for strong convective storms in central europe. *Atmos. Res.*, **161-162**, 153–168, https://doi.org/10.1016/j.atmosres.2015.03.020.

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

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

Kelly, D. L., J. T. Schaefer, R. T. McNulty, C. A. Doswell III, and R. F. Abbey Jr., 1978: An augmented tornado climatology. *Mon. Wea. Rev.*, **106**, 1172–1183, https://doi.org/10.1175/1520-0493(1978)106⟨1172:AATC⟩2.0.CO;2.

Klockow-McClain, K., and Coauthors, 2021: Nws severe warning philosophies across the conus: Implications for the implementation of facets. *101st Annual Meeting*, Virtual, Amer. Meteor. Soc., https://ams.confex.com/ams/101ANNUAL/prelim.cgi/Paper/384373.

Knol, M. J., W. R. Pestman, and D. E. Grobbee, 2011: The (mis)use of overlap of confidence intervals to assess effect modification. *Natl. Lib. Med.*, 253–254, https://doi.org/10.1007/s10654-011-9563-8.

Kucher, E. L., and P. M. D., 2006: Severe convective wind environments. *Wea. Forecasting*, **21**, 595–612, https://doi.org/10.1175/WAF931.1.

Loken, E. D., A. J. Clark, K. Calhoun, T. Sandmael, P. Skinner, P. C. Burke, P. Heinselman, and A. Reinhart, in review: Creating short-term severe weather hazard forecasts using probsevere, the warn-on-forecast system and random forests. *Wea. Forecasting*, accepted pending revisions.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1505–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

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

Loken, E. D., K. A. Wilson, T. Sandmael, A. J. Clark, K. M. Calhoun, A. E. Reinhart, P. Skinner, and P. C. Burke, 2022b: Improving probabilistic watch-to-warning severe hazard guidance by merging the warn-on-forecast system with observations-based products using machine learning. *30th Conf. on Severe Local Storms*, Santa Fe, NM, Amer. Meteor. Soc., https://ams.confex.com/ams/30SLS/meetingapp.cgi/Paper/407634.

Lundberg, S. M., and Coauthors, 2020: From local explanations to global understanding with explainable ai for trees. *Nat. Mach. Intell.*, **2**, 56–67, https://doi.org/10.1038/s42256-019-0138-9.

Lyza, A., and Coauthors, 2024: 2024 experimental warning program experiment summary. Tech. rep., NOAA Experimental Forecast Program of the Hazardous Weather Testbed.

Manabe, S., J. Smagorinsky, and R. F. Strickler, 1965: Simulated climatology of a general circulation model with a hydrologic cycle. *Mon. Wea. Rev.*, **93**, 769–798, https://doi.org/10.1175/1520-0493(1965)093⟨0769:SCOAGC⟩2.3.CO;2.

McGovern, A., R. J. Chase, M. Flora, D. J. Gagne II, R. Lagerquist, K. Potvin, C, N. Snook, and E. Loken, 2023: A review of machine learning for convective weather. *Artif. Intell. Earth Syst.*, **2**, https://doi.org/10.1175/AIES-D-22-0077.1.

McGovern, A., D. J. Gagne II, J. Basar, T. M. Hamill, and D. Margolin, 2015: Solar energy prediction: An international contest to initiate interdisciplinary research on compelling meteorological problems. *Bull. Amer. Meteor. Soc.*, **96**, 1388–1395, https://doi.org/10.1175/BAMS-D-14-00006.1.

Mitchell, E. D. W., S. V. Vasiloff, G. J. Stumpf, A. Witt, M. D. Eilts, J. T. Johnson, and K. W. Thomas, 1998: The national severe storms laboratory tornado detection algorithm. *Wea. Forecasting*, **13**, 353–366, https://doi.org/10.1175/1520-0434(1998)013⟨0352:TNSSLT⟩2.0.CO;2.

Molnar, C., 2025: *Interpretable Machine Learning*. 3rd. ed.

Monteleoni, C., G. A. Schmidt, and S. McQuade, 2013: Climate informatics: Accelerating discovering in climate science with machine learning. *IEEE Comput. Sci. Eng.*, **15**, 32–40, https://doi.org/10.1109/MCSE.2013.50.

Morss, R. E., J. K. Lazo, and J. L. Demuth, 2010: Examining the use of weather forecasts in decision scenarios: results from a us survey with implications for uncertainty communication. *Meteor. Appl.*, **17**, 149–162, https://doi.org/10.1002/met.196.

National Weather Service, 1982: National verification plan. Tech. rep., U.S. Dept. of Commerce, NOAA.

National Weather Service, 1986: Verification of severe local storm forecasts issued by the national severe storms forecast center. Tech. rep., U.S. Dept. of Commerce, NOAA.

Naylor, J., M. S. Gilmore, R. L. Thompson, R. Edwards, and R. B. Wilhelmson, 2012: Comparison of objective supercell identification techniques using an idealized

cloud model. *Mon. Wea. Rev.*, **140**, 2090–2102, https://doi.org/10.1175/MWR-D-11-00209.1.

NOAA, 2024: Storm events database. NOAA/National Centers for Environmental Information, accessed 2024, https://www.ncdc.noaa.gov/stormevents.

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

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

Rosenfeld, D., W. L. Woodley, A. Lerner, G. Kelman, and D. T. Lindsey, 2008: Satellite detection of severe convective storms by their retrieved vertical profiles of cloud particle effective radius and thermodynamic phase. *J. Geophys. Res.*, **113**, https://doi.org/10.1029/2007JD008600.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: Facets: A proposed next-generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Saabas, A., 2016: Random forest interpretation with scikit-learn. Accessed 2025, https://blog.datadive.net/random-forest-interpretation-with-scikit-learn/.

Sandmael, T. N., and Coauthors, 2023: The tornado probability algorithm: a probabilistic machine learning tornadic circulation detection algorithm. *Wea. Forecasting*, **38**, 445–466, https://doi.org/10.1175/WAF-D-22-0123.1.

Schmidt, T. G., and Coauthors, 2024: Gridded severe hail nowcasting using 3d u-nets, lightning observations, and the warn-on-forecast system. *Artif. Intell. Earth Syst.*, https://doi.org/10.1175/AIES-D-24-0026.1.

Schultz, C. J., W. A. Petersen, and L. D. Carey, 2009: Preliminary development and evaluation of lightning jump algorithms for the real-time detection of severe weather. *J. Appl. Meteor. Climatol.*, **48**, 2543–2563, https://doi.org/10.1175/2009JAMC2237.1.

Senkbeil, J. C., D. A. Scott, P. Guinazu-Walker, and M. S. Rockman, 2013: Ethnic and racial differences in tornado hazard perception, preparedness, and shelter lead time in tuscaloosa. *Prof. Geogr.*, **66**, 610–620, https://doi.org/10.1080/00330124.2013.826562.

Smith, B. T., T. E. Castellanos, A. C. Winters, C. M. Mead, A. R. Dean, and R. L. Thompson, 2013: Measured severe convective wind climatology and associated

convective modes of thunderstorms in the contiguous united states, 2003–09. *Wea. Forecasting*, **28**, 229–236, https://doi.org/10.1175/WAF-D-12-00096.1.

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

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. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

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

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system a vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

Storm Prediction Center, 2017: Spc products. NOAA, accessed 2024, https://www.spc.noaa.gov/misc/about.html.

Stumpf, G. J., T. M. Smith, K. L. Manross, and D. L. Andra Jr., 2008: The experimental warning program 2008 spring experiment at the noaa hazardous weather testbed. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., http://ams.confex.com/ams/24SLS/techprogram/paper_141712.htm.

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

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415, https://doi.org/10.1175/WAF925.1.

Wang, L., K. A. Scott, L. Xu, and D. A. Clausi, 2016: Sea ice concentration estimation during melt from dual-pol sar scenes using deep convolutional neural networks: A case study. *IEEE Trans. Geosci. Remote Sens.*, **54**, 4524 – 4533, https://doi.org/10.1109/TGRS.2016.2543660.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*, 3rd ed. International Geophysics Series, Vol. 100. Academic Press.

Williams, E., and Coauthors, 1999: The behavior of total lightning activity in severe florida thunderstorms. *Atmos. Res.*, **51**, 245–265, https://doi.org/10.1016/S0169-8095(99)00011-3.

Williams, J. K., 2013: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95**, 51–70, https://doi.org/10.1007/s10994-013-5346-7.

Williams, J. K., D. Ahijevych, S. Dettling, and M. Steiner, 2008: Combining observations and model data for short-term storm forecasting. *SPIE*, **7088**, 708 805, https://doi.org/10.1117/12.795737.

Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Mon. Wea. Rev.*, **147**, 2261–2282, https://doi.org/10.1175/MWR-D-18-0391.1.