UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

CORRECTING, IMPROVING, AND VERIFYING AUTOMATED GUIDANCE IN A NEW WARNING PARADIGM

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE IN METEOROLOGY

By

DAVID HARRISON
Norman, Oklahoma
2018

CORRECTING, IMPROVING, AND VERIFYING AUTOMATED GUIDANCE IN A NEW WARNING PARADIGM

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

Dr. Amy McGovern, Chair

Dr. Christopher Karstens, Co-Chair

Dr. Jeffrey Basara

Acknowledgements

This thesis would not have been possible without the help and guidance of my many friends and colleagues, and especially that of my co-advisers, Amy McGovern and Chris Karstens. Amy has guided my research since freshman year and unwaveringly supported my many endeavors through the intersection of computer science and meteorology. Chris was my first point of contact with the professional meteorological community, and has been an invaluable source of guidance and support for both my research and the development of my professional career. I cannot thank my co-advisers enough for all the help they have given me, and I look forward to working with them more in the future. I also wish to thank my entire committee for their time and suggestions, as well as Travis Smith and Alan Gerard for their additional guidance and insight. My fellow friends and members of the OU IDEA Lab were essential for feedback on this research and frequently provided me with new ideas to improve this thesis. In addition, I thank the many NWS forecasters, emergency managers, and broadcast meteorologists who participated in the PHI prototype experiments for their diligent efforts and insightful discussions, without which this research would not be possible. Finally, I must give special thanks to my parents, grandma, and uncle. I never could have made it this far without their abundant love and support.

This thesis was prepared with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

Table of Contents

List Of Tables

List Of Figures

Abstract

The prototype Probabilistic Hazards Information (PHI) system allows forecasters to experimentally issue dynamically evolving severe weather warning and advisory products in a testbed environment, providing hypothetical end users with specific probabilities that a given location will experience severe weather over a predicted time period. When issuing these products, forecasters are provided with an automated, first-guess storm identification object which is intended to support the probabilistic warning issuance process. However, empirical results from experimentation suggest forecasters have a general distrust of the automated guidance, leading to frequent adjustments to the automated information. Additionally, feedback from several years of experimentation suggest that forecasters have limited direct experience with how storm-scale severe weather probabilities tend to evolve in different convective situations.

To help address some of these concerns, the first part of this thesis provides a detailed analysis of the maximum attainable predictability of the automated guidance during the spring season of 2015, and offers a comparison of the verification statistics from automation to those of the corresponding storm-based warnings issued by the National Weather Service during the same time period. The second part of this thesis addresses storm-scale severe weather probability trends by developing a machine learning model to predict the evolution of a storm's likelihood of producing severe weather. This model uses the ensemble average of six machine learning members trained on variables obtained from the initial automated guidance, environmental parameters, and a storm's history, to predict future probabilities of severe weather occurrence over a predicted duration of a storm. Finally, the model was implemented and tested during the 2017 PHI prototype experiment.

Chapter 1

Introduction

Since the National Weather Service (NWS) first began issuing severe weather warnings and advisories in the mid-20th century, there has been little change in the way these warnings are communicated to the public. Arguably the greatest exception to this statement occurred in October 2007, when county-based warnings, which were primarily defined by county and state borders, were replaced operationally by storm-based warnings (SBWs; Ferree 2006, NOAA 2007). These SBWs gave NWS forecasters the ability to provide more geographically specific information about anticipated meteorological hazards through the use of polygonal warnings that may be, but are not always, independent from geopolitical boundaries (NWS 2009). The gradual introduction of impact-based warnings beginning in 2014 further provided NWS forecasters with the ability to provide specific details about the expected severity and impacts of a given severe weather threat (Losego et al. 2013; Harrison et al. 2014; Casteel 2016). However, these relatively new polygonal products still partially rely on similar methods used during the era of county-based warnings (e.g. county codes, teletype-style text, etc.) for dissemination to news media, emergency managers, and the general public (Coleman et al. 2011). Furthermore, much of the critical information offered by both county- and storm-based warnings, such as storm location and motion, is only updated every 15 to 20 minutes on average (Harrison and Karstens 2017) via severe weather statements (SVSs). This potentially

leaves end users to assess their personal risk with limited or old information, and may reduce the effectiveness of the current warning system (Drost et al. 2016).

In 2014, a vision for an alternative warning paradigm that partially addresses the aforementioned limitations, called Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2014), was introduced and tested in the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT; Karstens et al. 2015). Under this proposed new paradigm, NWS forecasters issue dynamically evolving, Probabilistic Hazards Information (PHI; Karstens et al. 2015, 2018, Hansen et al. 2017, Ling et al. 2017) warning plumes, which provide end users with specific probabilities that a given location will experience severe weather over a predicted time period.

The underlying goal of PHI is to enable NWS forecasters to provide end users with more detailed information more frequently than is currently available in the SBW paradigm. However, increasing the amount of information being produced necessarily increases the amount of time required to manually issue each product. This in turn reduces the frequency with which products can be issued and updated. To address this limitation, a partially automated system built around the the NOAA / Cooperative Institute for Meteorological Satellite Studies (CIMSS) Probability of Severe model (ProbSevere; Cintineo et al. 2014, 2018) was developed to handle some aspects of the warning issuance process, such as storm identification and tracking. Ideally, by removing these rudimentary but time consuming tasks from the warning issuance process, forecasters can cumulatively save time which can then be used to analyze data and maintain better situational awareness of the severe weather threat. In addition, the automated ProbSevere guidance provides predicted probabilities that

a given storm will be severe over a predicted time period, which forecasters can then use to produce the probabilistic warnings that make up the core of the PHI paradigm.

Although the automated system was intended to aid forecasters with the warning issuance process, several years of experimentation in the HWT have revealed that forecasters have a general distrust of the storm identification and tracking aspects of the automated guidance. This distrust has been shown to create a competitive dynamic between the forecasters and the automated system which is detrimental to the intended forecaster-guidance interdependence (Hoffman et al. 2017). Furthermore, forecasters have expressed a lack of direct experience with how storm-scale severe weather probabilities tend to evolve in different convective situations, resulting in reduced forecast reliability when issuing the probabilistic warnings (Karstens et al. 2015).

To address these issues and attempt to improve forecaster interdependence with the automated guidance, this thesis provides a detailed analysis of the maximum attainable predictability of the automated ProbSevere guidance during the spring season of 2015, and offers a comparison of the verification statistics from automation to those of the corresponding SBWs issued by the NWS during the same time period (Chapter 4). To facilitate this process, a new algorithm is developed to correct improper track breakages that occur in the ProbSevere model, which could negatively affect usability, verification, and predictability metrics (Chapter 3). The second part of this thesis addresses storm-scale severe weather probability trends by developing a machine learning model to predict the evolution of a storm's likelihood of producing severe weather (Chapter 5). Finally, the verification results and probability trend model were presented and tested during the 2017 PHI prototype experiment, and the results from that

experiment are provided (Chapter 6). A brief history and discussion of probabilistic forecasting, the ProbSevere model, and the PHI prototype warning system is provided in Chapter 2.

Chapter 2

Background

2.1 Probabilistic Forecasting

The concept of using probabilities to provide end-users with a measure of confidence in a weather forecast can be traced back as far as the late 18th century when J. Dalton (Dalton 1793) reportedly began issuing weather forecasts in the United Kingdom that included qualitative and occasionally quantitative expressions of uncertainty, such as, "The probability of rain was much smaller than at other times," (Sheynin 1984; Murphy 1998). Similar language has been noted by other pioneering meteorologists in the 1800s such as R. H. Scott (Scott 1869), who applied observations from the then-newly-established weather-observing networks in Europe and the United States to report explicit odds that a given weather feature would occur under specific observed conditions. By 1871, the United States Signal Service began issuing their first synoptic-scale weather forecasts, which occasionally included large-scale warnings for predicted storms. It is interesting to note that these warnings were initially referred to as "probabilities," even though there was no numerical probabilistic information included with the products (Murphy 1998).

Nichols (1890) is perhaps the first published example of using probabilistic forecasts as a means of improving a forecast's quality and value to the public (Murphy 1998). In particular, Nichols (1890) argued for the use of probabilistic

information to decide when to issue a forecast or warning for an event, stating, "Where the event is unimportant the probability should be relatively strong to justify predictions which may tend to discredit the reliability of the reports. But the greater the importance the smaller need be the probability involved, while to avoid the sacrifice of accuracy and confidence the problematic character of such predictions should, as far as possible, be indicated," (Nichols 1890; Murphy 1998). In other words, knowledge of how certain a forecast is increases the value of that forecast.

Most documented attempts at probabilistic forecasts prior to the 20th century were based on the subjective confidence of a forecaster (Murphy 1998). However, by the early 1900s, attention began to turn toward more objective probabilistic methods. For example, Besson (1904) attempted to develop a probabilistic model that could produce short-range forecasts of precipitation from observed atmospheric measurements through the use of contingency tables. This work was eventually furthered by Brier (1944, 1946), whose research ultimately culminated in the initiation of a nationwide operational precipitation probability forecast program in the United States in 1965 (Murphy 1998). The reader is referred to Murphy (1998) for additional details on the history of probabilistic forecasts prior to the 21st century.

Probabilities are now widely used in modern-day weather forecasts to characterize the likelihood of rain on a given day or the chance that severe weather will occur over a defined broad area. For example, the Storm Prediction Center (SPC) issues probabilistic forecasts daily that represent the likelihood of a severe storm report occurring within 25 miles of a given point location[1], and verification of these probabilistic forecasts has generally improved over recent

[1]http://www.spc.noaa.gov/misc/SPC_probotlk_info.html

years (Hitchens and Brooks 2012; Herman et al. 2018). Other studies, such as Gagne II et al. (2017) have focused on producing probabilistic models using machine learning techniques to forecast severe weather likelihood over large domains. While these probabilistic forecasts are generally valid for large spatial areas and long temporal periods, a number of studies have recently begun exploring the possible application of probabilistic information to storm-scale warning products. This concept of probabilistic severe convective storm warnings is founded on much of the historic probabilistic work described previously in this section. For example, Stumpf et al. (2015) noted that SBWs are dichotomous products which communicate that a location either will or will not see severe weather depending on if that location falls within the forecaster-generated warning polygon. Of course, severe weather rarely occurs at every point within a SBW, and this anecdotally results in many instances where the public might interpret the warning to be a false alarm, even if severe weather did occur somewhere within the polygon. Karstens et al. (2015) argues that this limitation of the current warning system may be partially overcome by providing probabilistic information representing a forecaster's confidence that a given location within a warning will experience severe weather. This concept particularly builds off of the work of Nichols (1890) by suggesting the inclusion of forecaster confidence could add value to the warning product. Following this line of study, Cintineo et al. (2014, 2018) have designed a statistical model which is capable of predicting the probability that a storm will produce severe weather over a given time period. Karstens et al. (2018) then utilize these diagnostic probabilities to generate experimental PHI warning products that include gridded probabilities representing the likelihood that a given location will experience severe weather over a predicted time period. These systems, known

as ProbSevere and the PHI prototype warning system respectively, provide the basis for this thesis and are discussed in detail in the following sections.

2.2 ProbSevere

The empirical NOAA/CIMSS ProbSevere model is designed to automatically extract information from a variety of live data sources and use that information to produce timely statistical forecasts of the likelihood that a storm will produce severe weather over the next 120 minutes (Cintineo et al. 2014, 2018). For the purposes of this research, the ProbSevere model can be broken down into two primary functions: storm identification and tracking, and probability prediction.

2.2.1 Storm Identification and Tracking

The storm identification and tracking aspect of the ProbSevere model primarily utilizes live Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) radar observations and derived products remapped onto a cylindrical equidistant projection with 0.01° latitude × 0.01° longitude spatial resolution (Cintineo et al. 2014, 2018). When first identifying potential storms, the ProbSevere model applies an enhanced watershed method (Lakshmanan et al. 2003) to a composite radar reflectivity (REF_{comp}) field in order to produce a radar-based storm object. In particular, the watershed method first searches for local maxima of $\text{REF}_{\text{comp}} \geq 35$ dBZ within a given domain. It then spatially grows an object outward from each local maxima in increments of 10 dBZ until the object is at least 20 km^2 in area. Once this spatial threshold is reached, the algorithm will return the final storm object and assign it an unique tracking identification number (Cintineo et al. 2014). Because this method utilizes centroid-based tracking techniques, large

changes in a storm's shape or size between radar scans may result in a significant shift of the associated object's centroid. When this occurs, the ProbSevere model may identify the storm object as a new storm and assign it a different tracking identification number, thus losing any information about that storm retained from previous time steps. This limitation of the ProbSevere model's tracking algorithm and an attempt to correct such breakages is discussed further in section 3.2.

The ProbSevere model also produces satellite-based storm objects by utilizing the convective cloud object tracking system developed by Sieglaff et al. (2013). Using this method, ProbSevere ingests satellite imager data from the Geostationary Operational Environmental Satellite (GOES; Menzel and Purdom 1994) network to calculate cloud-top phase (Pavolonis 2010a,b) and cloud emissivity ϵ_{tot} (Pavolonis 2010a). These variables are remapped to a cylindrical equidistant projection with 0.04° latitude × 0.04° longitude spatial resolution, which is then interpolated to the same 0.01° latitude × 0.01° longitude resolution of the radar data as described previously (Cintineo et al. 2014). ProbSevere then produces satellite-based storm objects by identifying collections of spatially connected pixels that contain a local maximum of ϵ_{tot}. As with the radar-based objects, a unique tracking identification number is assigned to each satellite-based object, and these objects are tracked spatially and temporally. Finally, the model calculates the temporal rates of change in ϵ_{tot} ($\Delta\epsilon_{tot}$) and the ice cloud fraction as determined from the cloud top phase (Δice, glaciation rate).

2.2.2 Probability Predictions

ProbSevere uses a naive Bayesian classifier (Kuncheva 2006; Domingos and Pazzani 1997) to determine the probability that an identified storm is currently or will be severe in the short term. As described in Cintineo et al. (2014), this classifier can be modeled given a set of observed predictors $\mathbf{F}$ such that

$$P(C_{sev}|\mathbf{F}) = \frac{P(C_{sev})P(\mathbf{F}|C_{sev})}{P(\mathbf{F})}, \tag{2.1}$$

where $P(C_{sev})$ is the probability that would be assigned if there were no predictors (i.e. the prior probability of severe).

Severe probability predictions are then made using variables derived from the radar- and satellite-based storm objects described previously as input observations, including a storm's maximum expected size of hail (MESH; Witt et al. 1998a), $\Delta\epsilon_{tot}$, and Δice. In addition, ProbSevere also utilizes select data fields from the Rapid Refresh numerical model (RAP; Benjamin et al. 2006), such as effective bulk shear (EBS; Thompson et al. 2007) and the most unstable CAPE (MUCAPE). These RAP data fields are first projected onto a cylindrical equidistant projection, and MUCAPE and EBS are calculated on every point in the remapped RAP grid for the current analysis (t_0), the previous hour analysis (t_{-1}), and for the next three hourly forecasts (t_1, t_2, t_3). The maximum value for each field at each grid point over the five forecast times is then is smoothed using a Gaussian kernel and smoothing radius equal to three standard deviations. Finally, the smoothed fields are converted onto a grid with 0.01° latitude × 0.01° longitude spatial resolution before being input into the statistical model. Therefore, all MRMS-, GOES-, and RAP-derived data are on the same cylindrical projection during evaluation of the Bayesian model (Cintineo et al. 2014). A list of all input variables and their sources is provided in Table 2.1.

Predictor name	Source
MUCAPE, EBS	RAP
Max expected size of hail (MESH)	MRMS
Max rate of change in the ϵ_{tot} in 2.5-h window ($\Delta\epsilon_{tot}$)	GOES
Max rate of change in the cloud-top ice fraction in 2.5-h window (Δice)	GOES

Table 2.1: ProbSevere model predictors and sources for 2014 - 2015. Modified from Cintineo et al. (2018), their Table 1.

Note that the probability of severe weather occurrence produced by the statistical model described above is a diagnostic prediction of severe weather. That is, the probability is a single value which identifies the total likelihood that a storm will be severe sometime during the next 120 minutes. This value does not indicate how the severe weather probabilities may change with time as the storm evolves. This limitation of the ProbSevere model and a new prognostic prediction method is the focus of Chapter 5.

2.3 Probabilistic Hazards Information

As mentioned in Chapter 1, PHI is an alternative warning paradigm intended to partially addresses some limitations of the current SBW system. Under this proposed new paradigm, NWS forecasters issue dynamically evolving, PHI warning plumes, which provide end users with specific probabilities that a given location will experience severe weather over a predicted time period.

When issuing the initial PHI products, forecasters are provided with a radar-based ProbSevere storm object, as well as the storm's diagnostic probability of producing severe weather. The current PHI prototype tool (Karstens et al.

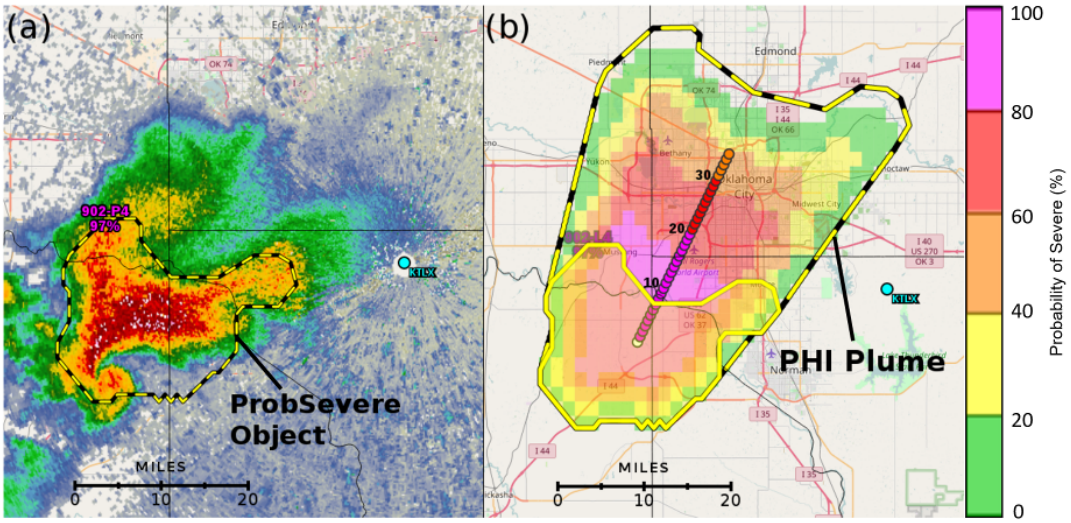

Figure 2.1: Example of a typical ProbSevere storm object (a) and its derived PHI warning plume (b).

2017, 2018) enables forecasters to then issue a warning plume representing the spatial area that would be swept out by the storm object over a given period of time (Fig. 2.1), allowing for some variance in storm motion. This is done by using the partially automated, diagnostic storm object and associated tracking algorithms produced by the ProbSevere model, or forecasters can manually draw their own diagnostic polygonal object around the storm or hazard. If an object and associated warning plume are produced manually, the forecaster is then responsible for frequently adjusting the geometric shape of the object as the storm evolves, and for ensuring the object and plume track correctly with the meteorological phenomenon. Conversely, if a forecaster uses the automated guidance from the ProbSevere model, these aspects of the warning are largely managed by the automated tracking algorithms, which the forecaster can override as necessary.

Unlike the current warning system, ProbSevere storm objects and derived PHI plumes propagate and evolve with the associated thunderstorms and are updated approximately every two minutes by either automated guidance or the forecaster. Additional products can then be derived from the PHI in order to provide information catered to individual end users, such as emergency managers and hospitals. The ultimate goal of the PHI paradigm is to provide end users with a continuous flow of specific information, ideally empowering them to more effectively assess their unique risk and take actions appropriate to their particular situation in the event of severe weather (Mileti and Sorensen 1990; Lindell and Perry 2012; NOAA 2011).

To help NWS forecasters transition more easily to the experimental PHI paradigm, the PHI warning issuance process in the current prototype tool was designed to be similar to that of SBWs, and many design elements of the user interface were drawn from the operational SBW generation software (WarnGEN). In order to issue a PHI warning plume using the automated ProbSevere guidance, forecasters first select a ProbSevere storm object to bring up the warning properties interface (Fig. 2.2). From this panel, forecasters can:

- Set the PHI product parameters by specifying if the product is a warning or advisory, identifying the maximum hail size and wind speed expected, and selecting a source for why those severe threats are expected (e.g. radar indicated, observed, etc.; Fig 2.2a).

- Adjust the geometry and motion vector of the initial ProbSevere object. Changes made to the ProbSevere object are also applied to the warning plume downstream (Fig. 2.2b). Changing any of the parameters in this

panel will automatically give the forecaster partial control over the resulting warning geometry, such that the forecaster is responsible for adjusting the changed variables in future warning updates.

- Specify how the forecaster's confidence of severe weather will change with time (Fig. 2.2c). The probability trend set in this panel is mapped to a grid underlying the warning plume polygon via a two-dimensional Gaussian distribution as described in Karstens et al. (2015, Fig. 2.2d). This process is discussed further in Chapters 5 and 6.

Once the warning properties are set, forecasters can then write a brief discussion of the warning and any information they believe to be relevant about the severe weather threat. After the warning is issued, this discussion is sent out to end-users along with the ProbSevere storm object and the PHI warning plume (Fig. 2.3). Forecasters are then able to update the warning as frequently as necessary simply by reselecting the associated ProbSevere storm object and modifying their previously issued warning properties as was done during the initial issuance process.

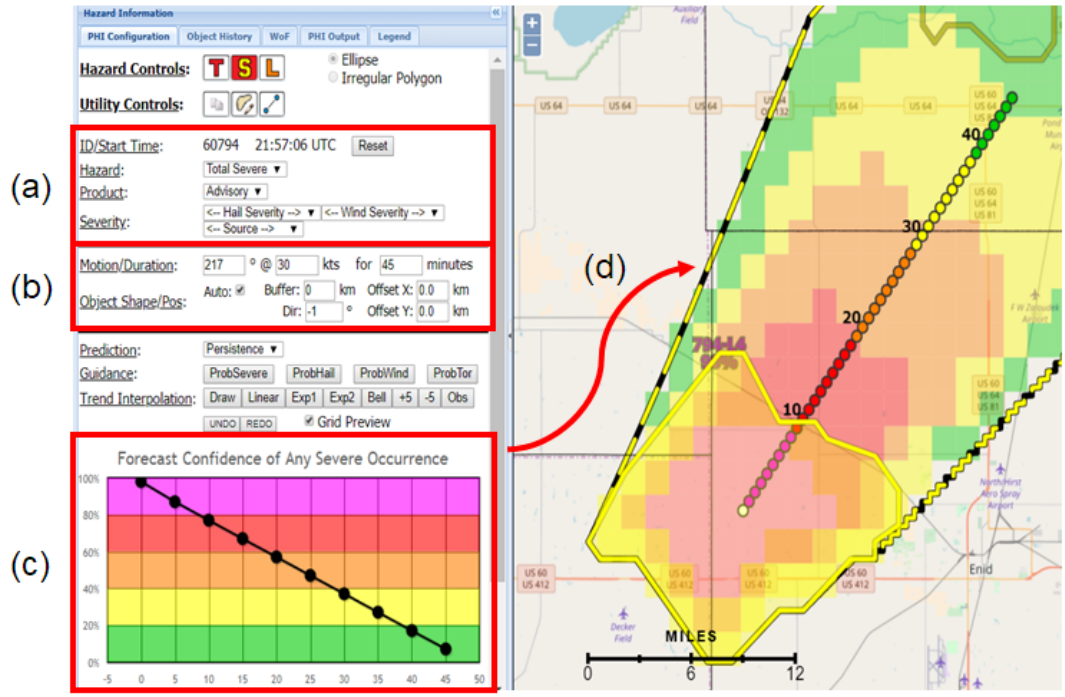

Figure 2.2: Example of the warning properties interface of the PHI prototype tool. Panel (a) contains the basic warning properties, (b) has information related to the warning's geometry and motion vector, and (c) is the forecast confidence trend which directly populates (d) the gridded probabilities of the PHI warning plume.

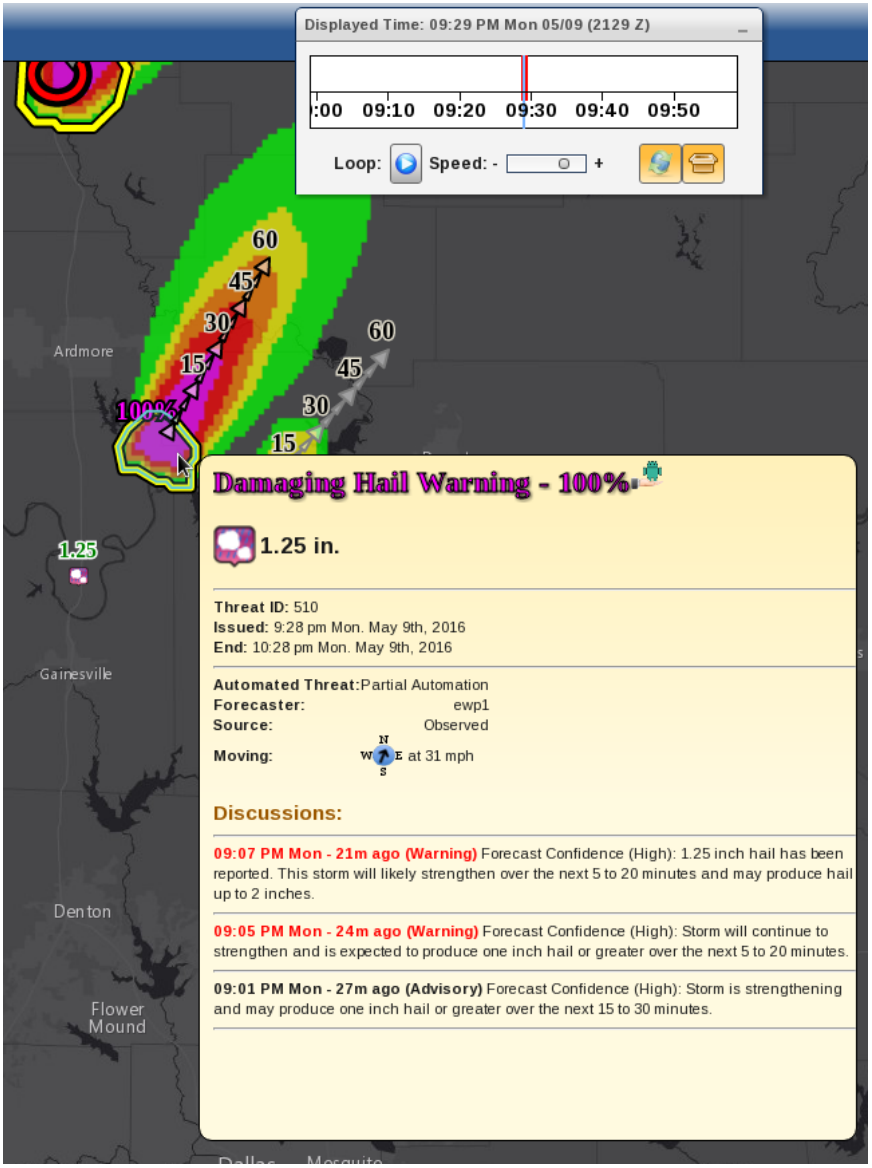

Figure 2.3: Example of the end-user display tool which displays any forecaster-warned ProbSevere storm objects, PHI warning plumes, and forecaster-issued discussions. From Karstens et al. (2018), their Fig. 8.

Chapter 3

Dataset

3.1 Data Selection

This thesis primarily focuses on identifying and augmenting the predictability and functionality of automated ProbSevere storm objects and their derived PHI warning plumes between 9 April 2015 and 30 June 2015. Due to the large size and density of the ProbSevere dataset, the spring and early summer months (April - June) of 2015 were selected as a quasi-representative subset of a typical severe weather season in the continental United States (CONUS). During this time period, 1,894,762 individual ProbSevere storm objects were archived internally by the National Severe Storms Laboratory, 13,135 unfiltered local storm reports (LSRs) were obtained from the Storm Prediction Center (SPC 2015), and 10,831 severe thunderstorm and tornado SBWs were collected from the NWS Performance Management database (NWS 2016).

3.2 Best Track: Real Time Data Corrections

During the 2015 and 2016 PHI prototype experiments, participating NWS forecasters noted that the automated guidance provided by the ProbSevere model would frequently lose tracking on certain storm objects, which would then cause

the forecaster-issued warnings and advisories to temporarily disappear and interrupt the flow of information to potential end users. In many of these instances, a storm object would undergo an apparent meteorological evolution, such as a merge or split with another storm, which would significantly alter the shape and/or size of the cell. When this occurred, the ProbSevere tracking usually continued to identify the storm as a tracked object, but would change the objects tracking identification number used to programmatically link the automated guidance to the forecaster-issued PHI products. Such tracking breakages that occur as a result of apparent meteorological processes are herein referred to as justified breakages. However, there were also many other instances during the experiments where the tracking identification number of a storm object would change in the absence of notable changes in the associated storm. These types of breakages are herein referred to as unjustified breakages, or breakages that occur unexpectedly. Regardless of the cause, forecasters often cited unreliable storm tracking as a factor in their decision to not use the automated guidance when issuing PHI warnings (Ling et al. 2017). Furthermore, from a research standpoint, storms that undergo unjustified tracking breakages may have reduced or incomplete lead-time statistics. Therefore, a correctional algorithm, named Best Track: Real Time (BTRT), was developed and implemented as part of this investigation.

3.2.1 Primary Comparison Algorithm

BTRT was written as a Python package designed to read in the output of a third-party storm identification and tracking algorithm and improve upon that algorithms tracking by correcting unjustified and some justified breaks in an objects

track. BTRT is loosely modeled after the Warning Decision Support Services-Integrated Information (WDSS-II; Lakshmanan et al. 2007) w2besttrack algorithm (Lakshmanan et al. 2015), but modified to support real-time processing of an operational models output. In particular, BTRT is what Lakshmanan et al. (2015) would describe as a causal implementation of the WDSS-II w2besttrack algorithm, as BTRT does not require knowledge about a storms complete (past and future) lifespan. Furthermore, the WDSS-II w2besttrack algorithm utilizes multiple iterations to systematically remove anomalous objects from first-guess storm tracks until the optimal track is obtained. BTRT accomplishes similar results in a single iteration by storing the history of each tracked object between runs and using it as context when comparing new objects to existing tracks as described below. This optimizes BTRT in terms of speed and real-time performance and makes it possible to perform the track corrections live in an operational environment.

Initially, the BTRT algorithm loads information about all tracked objects that were present prior to the current time step, including information about each objects centroid location, valid time, and tracking identification number (Fig. 3.1a). Once the history is known, the algorithm then evaluates any new objects that were produced by the ProbSevere model for the current time step. The identification number of each new storm object is compared to that of all objects from the previous time step, and any storm objects that have an equivalent identification number are automatically paired (Fig. 3.1b). The matched identification numbers are then retired such that no other new objects can be matched to those previously existing objects for the current time step. In doing so, BTRT assumes that the ProbSevere model has already correctly tracked those particular storms such that no corrections are needed. By the end of

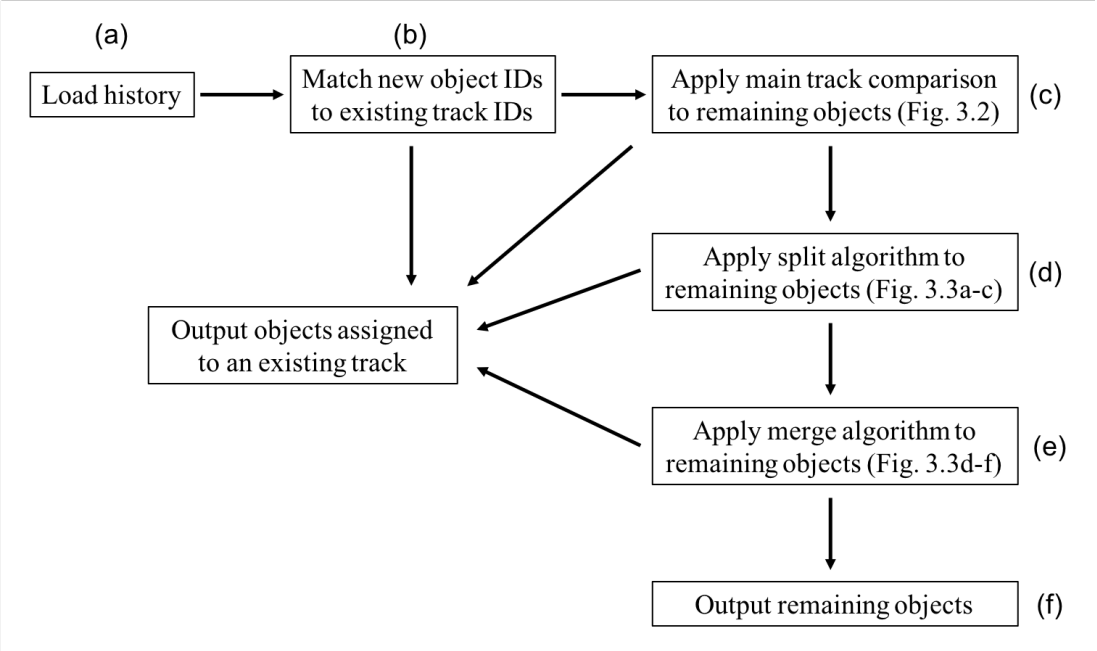

Figure 3.1: Flow chart of the BTRT algorithm.

this step, only new objects that are not already associated with an existing identification number should remain.

The main step of the BTRT algorithm (Fig. 3.1c) then employs an approach similar to that of the Storm Cell Identification and Tracking (SCIT) algorithm described in Johnson et al. (1998) in which it compares each remaining storm object at the current time step with every previously existing object an optimal match is found. First, BTRT generates a track, or collection of storm objects with the same identification number at different time steps, for each object in the previous time step. Then, the valid time of the last object in each track is compared with the time of each new object. If the last object of the track is older than a user-specified buffer time (3 minutes for this study), then the track is skipped. Otherwise, the new storm object and all previously existing objects within the track being compared are plotted using each objects centroid

as a point location as shown in Fig. 3.2a. BTRT then calculates the Theil-Sen fit (Theil 1950; Sen 1968; Lakshmanan et al. 2015) of the previously existing objects' centroid locations in order to derive a mean motion and speed for the physical phenomenon represented by the track (Fig. 3.2b). Next, the algorithm extrapolates the most recent storm object in the track forward in time along the Theil-Sen trajectory until the object is temporally co-located with the new object. If the centroid of the extrapolated object falls within a user-specified buffer distance (10 km) to the centroid of the new object (Fig. 3.2c), then the track is designated as a potential match and the algorithm moves on to the next track. If a new object is matched with multiple tracks, then the track with the smallest extrapolated distance to the new object is considered the best match, and the object's identification number is updated to match the objects contained within that track (Fig. 3.2d). The new object and all objects associated with the matched track are then removed from any additional processing for the current time step.

3.2.2 Merge/Split Algorithms

The main step of BTRT is intended to account for most unjustified track breakages that may occur in the ProbSevere model; however, it may not correctly identify justified breaks. For example, when large or elongated objects split or merge together, there is a large spatial shift in the objects centroids, and the resulting track breakages may not be identified using the primary track comparison method. This is one disadvantage to utilizing a centroid-based tracking method, as the location of a mass-weighted centroid is sensitive to changes in an objects shape and size. To partially account for this limitation, any remaining

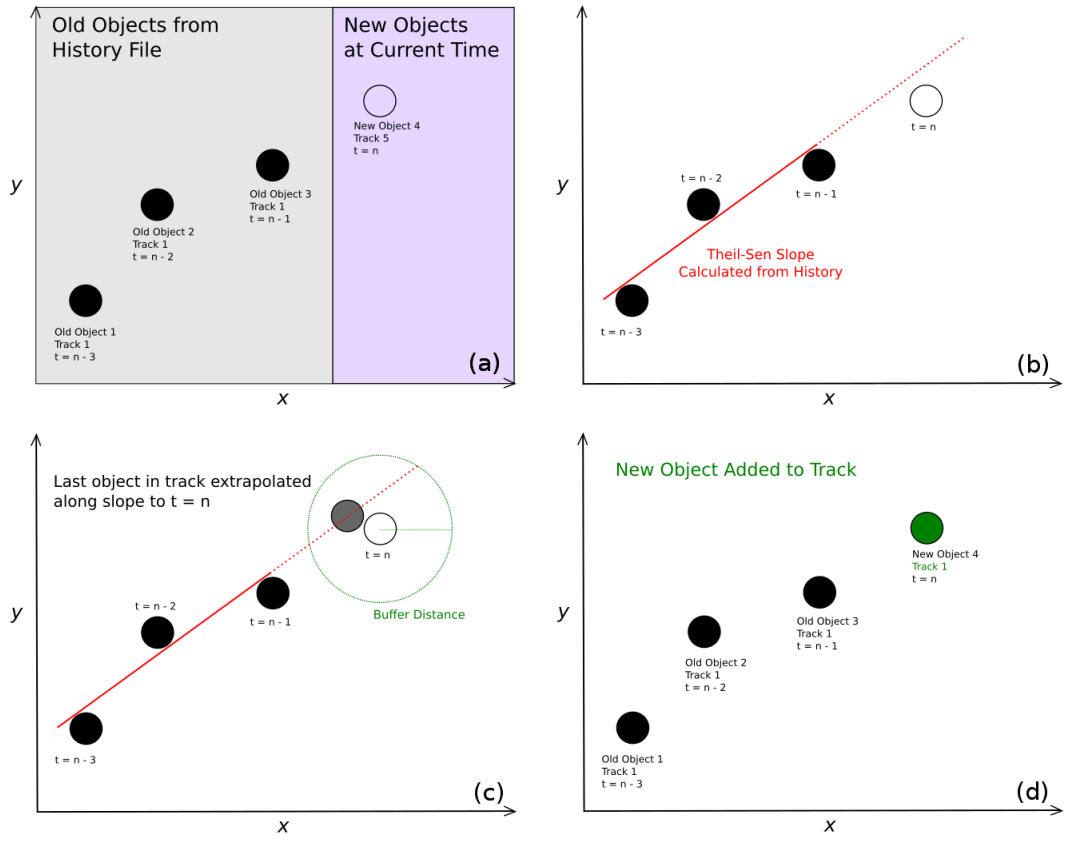

Figure 3.2: Schematic demonstrating the main track comparison step of BTRT. (a) Objects at previous and current time steps are spatially plotted as point locations. (b) The Theil-Sen fit is calculated for the old objects within a track. (c) The last old object in the track is extrapolated along the Theil-Sen fit to the current time step. Its new position is then compared with the current object. (d) If the extrapolated object falls within a specified buffer distance to the center of the current object, then the current object is added to that track.

objects that weren't matched with a track are then processed using BTRT split and merge algorithms (Fig. 3.1d,e).

When a large object splits, two or more objects often result that have different identification numbers than the original (Fig. 3.3a). To solve this, the BTRT split algorithm utilizes a similar process to the main track comparison, where each remaining new object is compared to each remaining potential track (Fig. 3.1d). As before, the Theil-Sen trajectory for the track is calculated and the most recent storm object in that track is extrapolated forward to the current time. However, this time a static 5 km buffer is added around the perimeter of the extrapolated object (Fig. 3.3b) to account for more varied motions and evolutions. (Note that this buffer and all other user-defined values chosen in this study were selected by anecdotally reviewing a number of case studies and comparing the results to those of WDSS-II w2besttrack.) If the centroid of the new object is contained within the buffered, extrapolated object, then the new object is added to that track, and the object and track are both removed from further processing (Fig. 3.3c). If multiple new objects are contained within the extrapolated object, BTRT will prioritize the objects by their ProbSevere probabilities such that the object with the greatest probability prediction at the current time step is updated to match the identification number of the objects within that track. The current PHI prototype tool requires that no two objects have the same identification number at a given time. Therefore, only one of the objects resulting from a split may retain the original identification number.

A similar process is used to handle merge operations (Fig. 3.1e), where two or more storm objects are combined into a single storm object which may have a different identification number than any of the original objects (Fig. 3.3d). Again, the Theil-Sen trajectory for each potential track is calculated and the

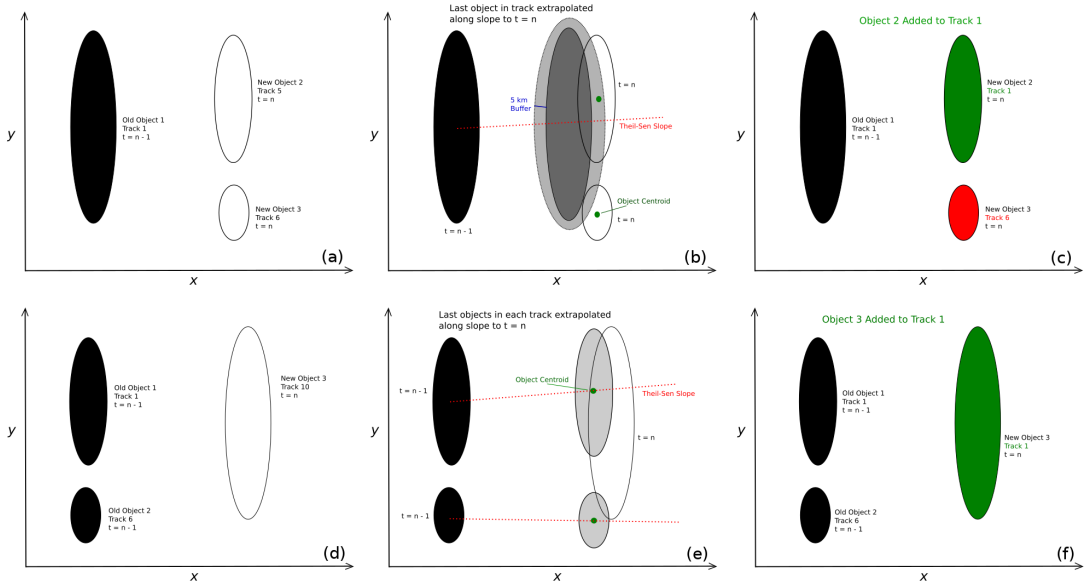

Figure 3.3: Schematic showing the split and merge algorithms of BTRT. (a) The track breaks when one large object splits into two or more. (b) The original object is extrapolated along the Theil-Sen fit to the current time step and a 5km buffer is applied. (c) If the centroid of one of the current objects is contained within the buffered, extrapolated object, then that current object is added back to the original track. (d) The Track breaks when two smaller objects merge into one large object. (e) The previous objects are extrapolated along their Theil-Sen fits to the current time step. (f) If the centroid of one of the original objects falls within the current object, the current object is added to that track.

last objects in each track are extrapolated forward to the current time. This time, a 5 km static buffer is applied to the perimeter of the new storm object. If the centroid of the extrapolated object is contained within the buffered new object (Fig. 3.3e), then the new object is updated to have the same identification number as the objects in that track (Fig. 3.3f). As before, if multiple storm objects are contained within the new object, BTRT will prioritize objects based on the ProbSevere probability prediction. Because two or more tracks are being merged into one, only one of the identification numbers is preserved. Note that while the split and merge algorithms account for some variation in an objects centroid location, they may still be limited by the aforementioned sensitivities of centroid-based tracking methods. Additional improvement may be possible by implementing an algorithm that considers overlap in an objects area between time steps; however, this method was not tested in this study. Additional documentation of the BTRT algorithm along with the operational, open-source Python code can be found online[1].

3.2.3 Application

The BTRT algorithm was applied to the ProbSevere dataset at the beginning of this study, resulting in 1,520,330 corrected ProbSevere storm objects with unique identification numbers, or about a 20% reduction from the 1,894,762 original tracked objects. For reference, a Python implementation of the WDSS-II w2besttrack algorithm was also run on the dataset using similar settings, which resulted in 767,247 corrected tracks, or a 60% reduction. Note, however, that the w2besttrack algorithm implements additional thresholds which act as a noise filter, such that a track must exist for at least three time steps to remain in

[1] https://github.com/arkweather/BestTrackRT

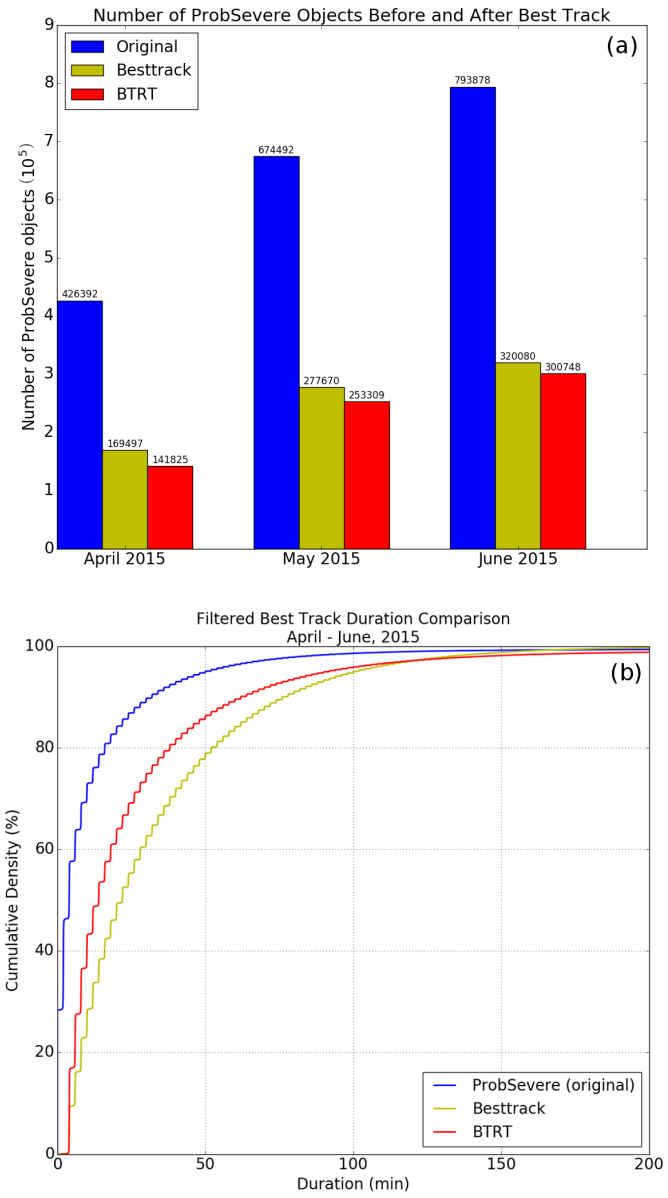

Figure 3.4: (a) Monthly breakdown of the number of ProbSevere objects remaining in the dataset after no track corrections (blue), filtered track corrections by the BTRT algorithm (red), and track corrections by a Python adaptation of the WDSS-II w2besttrack algorithm (yellow) are applied. (b) Cumulative track duration density after no corrections (blue), filtered BTRT (red), and WDSS-II w2besttrack (yellow) are applied to the entire dataset (April June 2015).

the final output. To account for this discrepancy, a similar filter was temporarily applied to the BTRT algorithm, resulting in 695,882 corrected tracks, or a 63% reduction (Fig. 3.4a). The filtered BTRT algorithm and w2besttrack also produced comparable increases in total track durations as demonstrated by the cumulative density functions shown in Fig. 3.4b. These statistics show that BTRT is able to successfully correct the ProbSevere tracking at a rate comparable to the standard WDSS-II w2besttrack algorithm in real time.

Chapter 4

Conditional ProbSevere Verification

PHI prototype experiments held in the HWT during the springs of 2014, 2015, and 2016 consistently demonstrated that the creation of a PHI warning plume from a manual storm object required notably more time to perform compared to the same task using the automated guidance. For instance, during the 2016 PHI prototype experiment, Karstens et al. (2017) determined that a PHI warning on average required a full minute longer to issue when done manually than when using the automated guidance. Subsequent updates to these warning plumes exhibited similar results, with an average of about 45 extra seconds needed to update plumes derived from manual objects. In addition, Karstens et al. (2015) found that forecasters incurred workload limitations while maintaining more than four or five manual objects at a time. However, during formal and informal discussions, many of the NWS forecasters that participated in the experiments expressed a distrust of the storm identification and tracking aspects of the automated ProbSevere guidance, and more than one-third of all new PHI plumes were created from manual storm objects during the 2016 experiment.

I hypothesize that much of this distrust in the automation stems from a combination of inherent, systematic technical considerations, such as unjustified storm tracking breakages, and a lack of experience, understanding, and/or proven verification with the automated guidance. This hypothesis is supported

by discussion presented in Roebber et al. (2002), which asserts that forecasters require information about the performance characteristics of a model before that model can be used effectively. To improve forecaster trust and better establish forecaster interdependence with the automated PHI guidance, it is then relevant to assess the predictability of the ProbSevere model as applied to the FACETs/PHI paradigm [i.e. how probability of detection (POD) changes with increasing lead time], and to identify potential limitations of the system (Saha and van den Dool 1988; DelSole 2004). Therefore, the first goal of this thesis is to provide a detailed analysis of the maximum attainable predictability of the automated ProbSevere guidance during the spring of 2015, and compare the automation's conditional POD statistics to those of the corresponding SBWs issued by the NWS during the same time period.

4.1 Plume Verification Methods

Recall from section 2.2 that ProbSevere objects are automatically generated for every identifiable storm within a domain that contains a composite radar reflectivity maximum $\geq$ 35 dBZ and spatially covers an area of at least 20 km^2 as specified in Cintineo et al. (2014). Of course, not every storm with a ProbSevere object produces or is expected to produce severe weather, but rather human forecasters are required to apply their meteorological knowledge to identify which storms are deserving of a warning. As stated previously, this chapter is intended to identify the maximum attainable POD of the automated ProbSevere guidance and derived PHI plumes in an operational environment. To accomplish this, a number of experimental assumptions were made to simulate forecaster input in the warning process:

1. The ProbSevere guidance was implemented as it existed in 2015 but with the corrections from BTRT applied to the automated storm objects.

2. Hypothetical forecasters issued PHI warnings only on storms that produced at least one LSR and on every storm that produced at least one LSR.

3. Hypothetical forecasters used only the ProbSevere guidance for storm identification and tracking and did not create any manual storm objects or override the automated guidance settings.

4. PHI warning durations and valid times were provided by machine learning guidance described in McGovern et al. (2017, 2018b), and durations did not exceed 120 minutes.

5. PHI warnings were the same shape and size of the derived PHI plumes.

6. Only PHI plumes with a forecast duration that included the time of an LSR were considered to be warnings.

These experimental assumptions can be stated more simply as, "The hypothetical forecasters always issued warnings that contained or were very near an LSR at the earliest possible lead time." However, there are a number of limitations with this method. For example, by assuming the hypothetical forecasters only issue PHI warnings on storms that produce an LSR, false positives can only occur in this analysis if the automated ProbSevere geometries or storm motion calculations result in a warning plume that doesnt cover the point LSR location. While this is necessary to determine the maximum attainable POD of the automated guidance, it also artificially biases the false alarm rate (FAR). To avoid this limitation, a more detailed analysis of unconditional FAR and

false alarm area (FAA; Karstens et al. 2015; Stumpf et al. 2015) was not considered in this analysis, though future research may be able to use grid-based techniques such as those described in Stumpf et al. (2015) to compute these metrics. Furthermore, warnings were chosen to be the same size as the derived PHI plumes to stay consistent with methods used during the 2017 PHI prototype experiment (Karstens et al. 2018). This was also necessary to determine the maximum attainable POD for a given warning plume, as defining warnings based on some probability threshold would necessarily reduce the size of the warning plume and may decrease that warnings POD and lead time (Karstens et al. 2015). For this reason Severe Weather Statements (SVSs) were also excluded when considering SBWs, as they only act to reduce the size of a SBW and therefore cannot increase the POD or lead time provided by that warning.

Before verification metrics could be calculated, it was first necessary to determine which tracked thunderstorm most likely produced each LSR. To accomplish this, all automated ProbSevere storm objects valid at the time of a given LSR were compared, and any object geometry that geospatially contained the LSR was matched with that LSR (Fig. 4.1a). If the LSR was not directly contained within a storm object at the time of the report, then the nearest object to the reports point location (within 20 miles; 33 km) was considered to be a relative match to allow for error in the reports time and location (Fig. 4.1b). In this system, a single tracked storm could be associated with multiple LSRs, but an LSR could only be assigned to a single storm. ProbSevere storm objects that werent matched to an LSR were removed from the rest of the thesis. Similarly, LSRs were also assigned to NWS-issued SBWs by determining if a warning polygon spatially contained an LSR during its valid duration. If more than one warning contained an LSR, the warning which provided the most lead

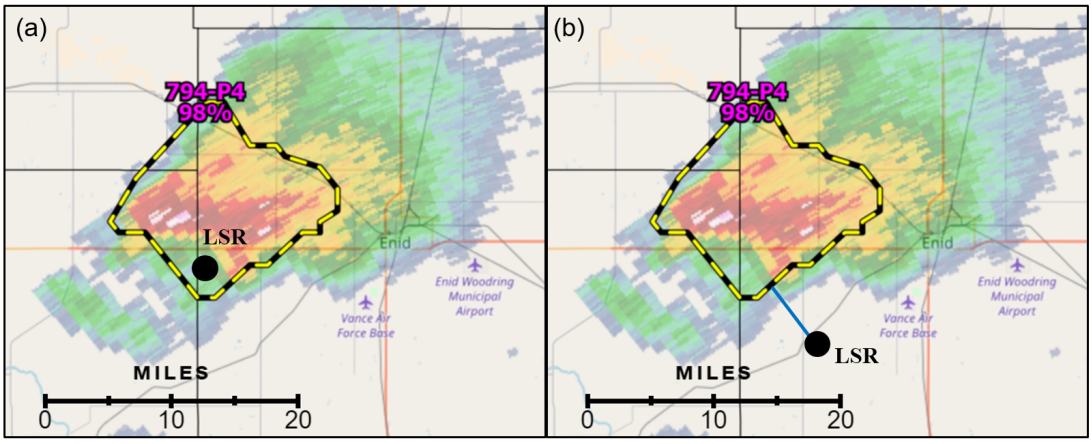

Figure 4.1: Example of how ProbSevere storm objects were match with LSRs. An LSR was matched to a ProbSevere object if (a) the object contained the LSR at the time of occurrence, or (b) the outer polygon of the storm object was the closest spatially to the LSR and less than 20 miles (33 km) from the LSR at the time of occurrence.

time was matched to the LSR. Any unwarned LSRs were then matched with the closest warning, and all remaining unmatched SBWs were removed.

To account for the rapid update frequency of automated ProbSevere objects, it was also necessary to consider objects and derived PHI plumes that were valid prior to the time of an LSR. Once each LSR was matched with a ProbSevere storm object, derived PHI plumes were analyzed for each time step in that object's history. Any individual plumes within the object's history that both contained the LSR and had a forecast duration that included the time of the LSR were classified as "hits," which provided lead time for the location impacted by the hazard (Fig. 4.2a). Similarly, plumes valid for the time of the LSR that did not contain the LSR were classified as "misses" (Fig. 4.2b). This method was applied to the NWS-issued SBWs as well, such that any warnings matched to an LSR that contained that LSR were classified as "hits," and any warnings

associated with an LSR that did not spatially contain that LSR were classified as "misses". The lead time provided by both the PHI plumes and the SBWs was then calculated as the time difference between the LSR and the issue time of the product. Finally, all classifications were binned into 5-minute lead time intervals and used to assess the predictability of the automated objects and SBWs conditioned by the experimental assumptions stated previously. Caution is advised when interpreting the results presented herein, as manually generated SBWs and the automated PHI warning plumes in this thesis represent very different paradigms that ultimately limit the comparability of the two systems. The methods used in this analysis attempt to make this comparison as equitable as possible for the sake of garnering trust with the forecasters who would have to bridge the gap between these two paradigms.

4.2 Plume Verification Results and Statistics

4.2.1 Conditional Verification

Out of the 1,520,330 BTRT-corrected, automated ProbSevere objects analyzed in this thesis, only 4,937 objects ($< 1\%$) were associated with a storm that produced at least one LSR. From these, a total of 617,868 PHI warning plumes were derived using 2-minute intervals corresponding to the update times of the automated ProbSevere storm objects. Similarly, 4,808 SBWs of the 10,831 analyzed (44%) were associated with at least one LSR.

About 81% of all PHI warning plumes generated 0 - 10 minutes prior to the time of an LSR contained the location impacted by that LSR (0 - 10-min lead time; Fig. 4.3), including about 84% of warning plumes verified by hail reports, 75% of those verified by tornado reports, and 80% of warning plumes verified by

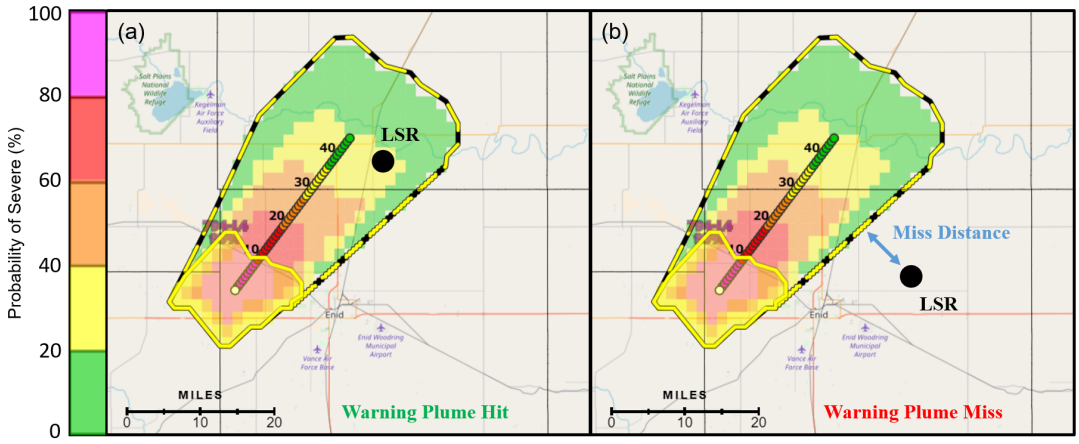

Figure 4.2: Example of how PHI warning plumes were classified as a "hit" or a "miss." (a) A plume was classified as a "hit" if it contained the LSR and had a duration that included the time of the LSR. (b) A plume was classified as a "miss" if it did not contain the LSR and had a duration that included the time of the LSR. A plumes miss distance is the shortest distance between the LSR and the outer polygon of the warning plume.

damaging wind reports. Notably, approximately 28% of conditionally verified PHI warning plumes provided around 60 minutes of lead time for a location impacted by an LSR, and nearly 10% of warning plumes verified by a tornado report provided 80 minutes of lead time.

These results show the extrapolative skill of the automated ProbSevere model and associated algorithms for detecting both the location and timing of severe weather occurrence as a function of lead-time conditioned by the experimental assumptions stated in section 4.1. Alternatively, one could consider this as a conditional measure of the success ratio, or 1 - conditional FAR. For example, this forecast-centric perspective specifically identifies the fraction of PHI warning plumes that were both valid at the time of an LSR and also contained that LSR at a given lead time. Therefore, the conditional FAR for these warning plumes would simply be the fraction of valid plumes which did not contain an LSR at a given lead time, or the inverse of the plot shown in Fig. 4.3). For instance, the warning plumes exhibited a conditional FAR of 19% at a lead time of 0 minutes, or $1 - 0.81$. In either case, the decreasing number of warning plume "hits" with increasing lead time demonstrated by this technique is consistent with prior conceptual work (Brooks et al. 1992) and applications (Hwang et al. 2015). Note, however, that the fraction of warning plumes verified by an LSR also exhibited little increase between 0 and 10 minutes of lead time, such that warning plumes issued a minute or two prior to the time of an LSR were not necessarily more likely to contain the LSR than a plume issued with 5 to 10 minutes of lead time. This signal may approximate a limit of predictability of the current ProbSevere guidance, where around 20% of PHI warning plumes issued for a storm that produces an LSR miss that LSR regardless of lead time.

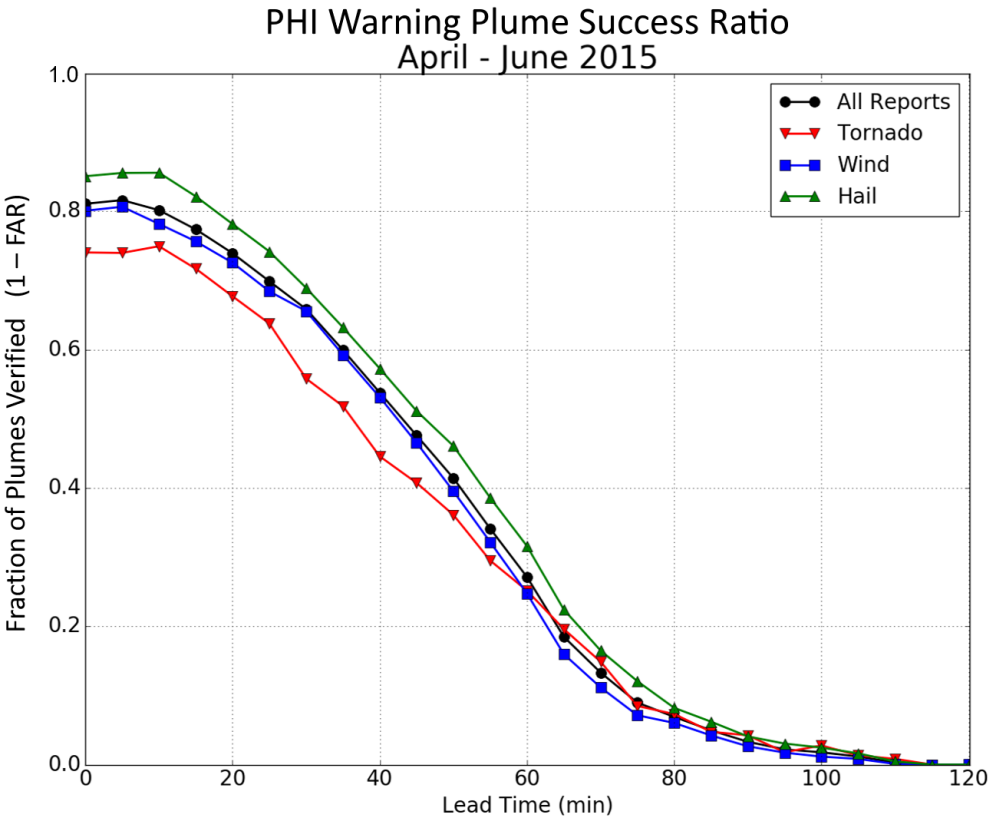

Figure 4.3: The fraction of plumes that contained an LSR at a given lead time. Alternatively, this is equivalent to the conditional success ratio (1 - conditional FAR), or one minus the fraction of plumes that did not contain an LSR at a given lead time.

To further investigate this potential predictability limitation of the ProbSevere guidance, it was necessary to determine how PHI warning plumes compare to operational SBWs. PHI plumes anecdotally benefit from the ProbSevere models ability to track a storm via frequent updates and automated tracking algorithms (Karstens et al. 2015). However, since SBWs are static products that cannot be moved once issued (except to reduce the warning area via an SVS), an observation-centric perspective is needed to generally compare the conditional PHI products with operational SBWs. To accomplish this task, the maximum lead time of each LSR was determined by first identifying the earliest PHI warning plume or SBW to contain the LSR, then calculating the difference in that products issue time and the time of the LSR. Fig. 4.4 then shows the cumulative fraction of LSRs that had a maximum lead time greater than or equal to a given lead time. Using this method, traditional POD metrics can be identified by analyzing an instantaneous lead time on the graph. Furthermore, by considering only the first PHI warning plume to contain the LSR and treating it as a static warning polygon, it is possible to provide a general comparison of the automated guidance to the operational SBWs.

PHI warning plumes exhibited a cumulative conditional POD of about 80% at the time a report was received (Fig. 4.4), with a cumulative POD of about 15% at a lead time of 60 minutes prior to a report. Notably, the POD of warning plumes increased to about 18% at 60 minutes of lead time when only considering areas impacted by either hail or a reported tornado.

SBWs generally exhibited similar POD to the automated PHI plumes, with about 78% of all LSRs warned at or prior to the time of occurrence (Fig. 4.4). Directly overlaying these results with the equivalent performance statistics for the conditional PHI warning plumes (Fig. 4.4) reveals that SBWs performed

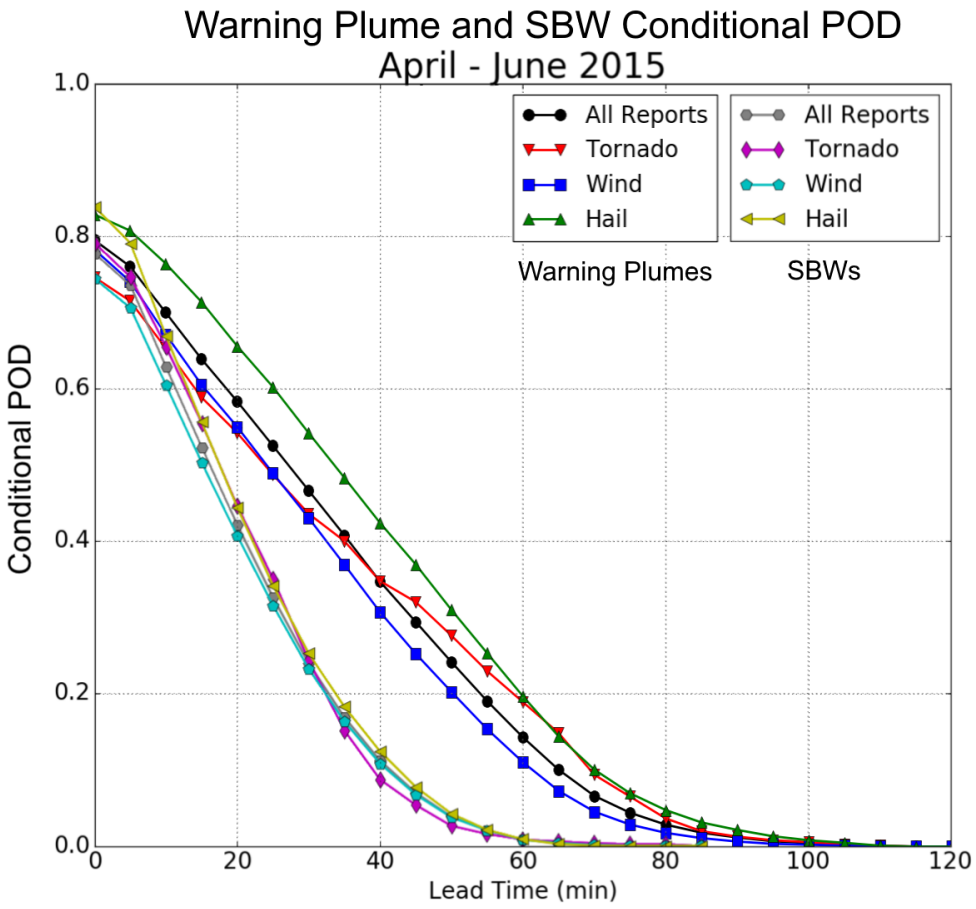

Figure 4.4: Comparison of the cumulative conditional POD of PHI warning plumes and SBWS. The maximum lead time for PHI plumes was determined by the issue time of the first plume to contain the warning.

remarkably consistently for all severe weather hazards over all lead times, such that no one hazard type had a notably higher cumulative POD than the others. In contrast, the automated guidance generally performed better with hail reports at all lead times, and by both hail and tornado reports at longer lead times. These results are perhaps not surprising, as the NOAA/CIMSS ProbSevere algorithm uses merged composite radar reflectivity and derived MESH as inputs when identifying storm cells (Cintineo et al. 2014). Tornadoes and damaging wind gusts, however, conceptually can and perhaps often occur on the outer flanks of a thunderstorm, where radar reflectivity values may not be as high. As a result, tornado and wind reports may not be well covered by ProbSevere objects and derived warning plumes at the time of the report, but should generally still fall within a PHI plume prior to the severe weather occurrence. These findings again demonstrate a potential limit of predictability resulting from the precision of the ProbSevere model, and may partially explain the lack of improvement in the number of warning plume "hits" at shorter lead times noted previously. These potential limitations are discussed further in section 4.2.2.

The automated PHI exhibited similar POD to the NWS-issued SBWs up to about 10 minutes of lead time. Beyond 10 minutes, the fraction of LSRs warned decayed at a much steeper rate for SBWs than for PHI, particularly in the 20- to 40-minute timeframe. At these lead times, 15% to 20% more LSRs were warned by a PHI warning plume than by an SBW, and 20% to 25% more hail reports in particular. One likely explanation for these discrepancies can be found in the official policy governing the issuance of SBWs. During 2015, NWS policy (NWS 2014) recommended that tornado SBWs should generally have a maximum valid duration of 45 minutes, while severe thunderstorm SBWs should

typically have a duration of 60 minutes or less. Harrison and Karstens (2017) found that only 5% and 1% of tornado and severe thunderstorm SBWs respectively have exceeded these directives since the operational implementation of the SBW paradigm in 2007, and a majority of warnings had durations much shorter than the maximum recommended. However, automated PHI warning plumes are currently restricted to a maximum duration of 120 minutes, and these durations are determined at each issuance using a predictive machine-learning technique (McGovern et al. 2017, 2018b). These potentially longer durations likely account for much of the increased POD of the PHI plumes at longer lead times compared to the SBWs. Recall, however, that the unconditioned FAR and FAA of the automated ProbSevere objects and derived PHI products are not considered in this thesis. As mentioned before, less than 1% of all automated ProbSevere objects were associated with a storm that produced an LSR. Therefore, in this alternative warning paradigm, forecasters would still be required to apply their meteorological knowledge and experience to determine which automated objects should be assigned a warning, similar to how SBWs are issued today. Specifically, these results suggest that NWS forecasters would be able to maintain or improve upon current predictability and lead time metrics while providing end-users with frequent probabilistic information simply by making the same warning decisions with PHI as they currently make in the SBW paradigm.

In addition to the predictability assessment, it is worth investigating PHI warning plumes that were issued for a storm that produced an LSR but did not actually contain that LSR. When the PHI plumes missed an LSR at short lead times, they missed by a short distance, and this distance generally increased with increasing lead time. PHI warning plumes missed the impacted area by

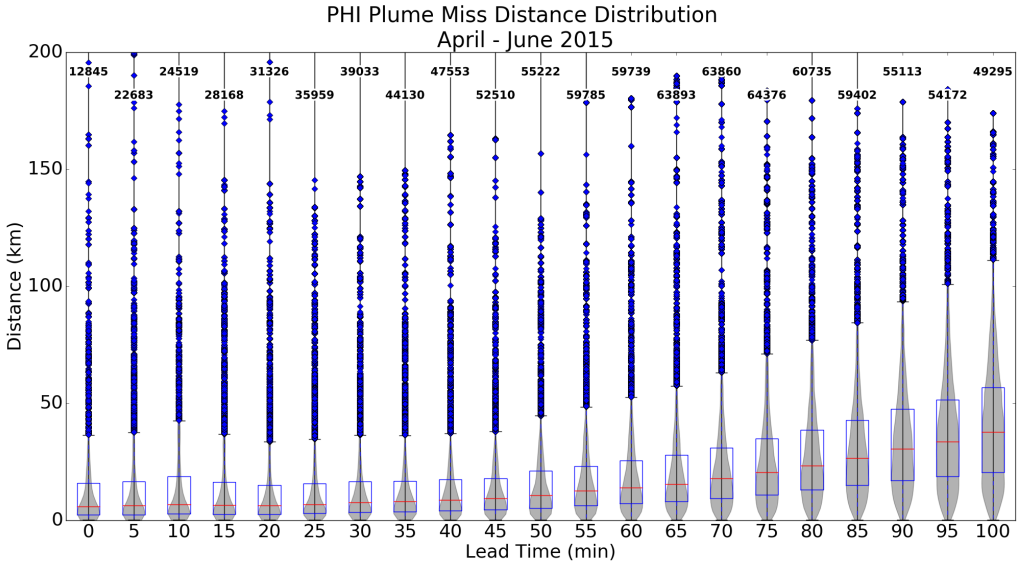

Figure 4.5: Miss distance distributions [violin plots (Hintz and Nelson 1998)] for PHI warning plumes that did not contain an LSR as a function of lead time. The number of plumes in each bin is listed above each violin.

an average of 10 km at the time of the LSR, with a median miss distance of about 5 km (Fig. 4.5). Plumes that missed an LSR at 10 minutes of lead time missed by an average of 13 km, and this increased to about 21 km for plumes valid 60 minutes before the LSR was received. It should be noted that numerous studies (e.g. Witt et al. 1998b; Trapp et al. 2006; Elsner et al. 2013) have discussed the limitations of using storm reports for warning verification, including errors in report times and locations. However, anecdotal case studies and results obtained during the 2015 and 2016 PHI prototype experiments also suggest that the automated ProbSevere objects may occasionally be too precise in their depictions of severe weather hazards, potentially missing severe events that occur around and beyond the outer flanks of a storm. These results introduced questions about how the PHI systems conditional predictability could be enhanced if the automated ProbSevere objects and derived PHI plumes were

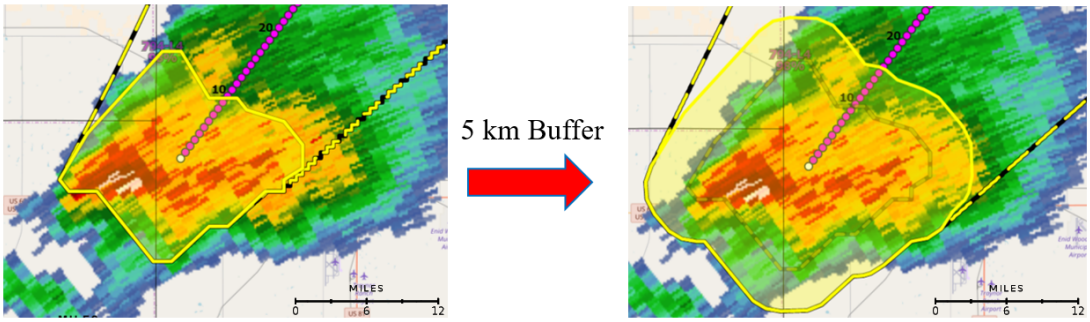

Figure 4.6: Example of a 5 km buffer applied to a typical ProbSevere object.

expanded spatially to cover a larger area. Therefore, additional research was performed to assess how the predictability of spatially buffered ProbSevere objects compares to that of unbuffered objects, particularly for severe events that may occur beyond the borders of the original first-guess guidance.

4.2.2 Buffered Verification

To determine how increasing the size of an automated ProbSevere object would impact predictability, a series of buffers were applied to each PHI plume in this analysis, such that the outer edge of each plume was systematically expanded outward from the centroid by 2, 5, and 10 km (Fig. 4.6). These modifications were first manually performed on a subset of plumes, and then synthetically applied to the rest of the dataset by reclassifying any plumes that missed a report by less than 2, 5, or 10 km respectively as a "hit". The synthetic method was then checked against the manually buffered objects to verify the accuracy of the results.

Under the forecast-centric perspective, the percent of plumes verified (success ratio) at the time of a report saw a notable increase with just a 2 km buffer applied, increasing from about 80% (Fig. 4.7a) to 84% (Fig. 4.7b) of warning

plumes verified when considering all reports. Applying 5 km (Fig. 4.7c) and 10 km (Fig. 4.7d) buffers to the automated warning plumes resulted in additional improvement, with about 89% and 93% of plumes containing an LSR at the time of the report, respectively. Similar improvements were also noted at longer lead times, particularly in the percent of plumes verified between 20 and 40 minutes prior to the report. As might be expected, these results suggest that applying geometric buffers to the original ProbSevere objects can systematically reduce the conditional FAR of the automated guidance by partially accounting for uncertainty in a severe storms location and error in the location of the associated LSR. However, the reduced improvement in the number of warning plume "hits" at shorter lead times discussed in section 4.2.1 not only remained, but extended to longer lead times with increasing buffer size. For instance, the percent of plumes verified by an LSR began to exhibit reduced improvement at around 10 minutes of lead time with a 2 km buffer applied. With a 5 km buffer, this threshold extended to about 20 minutes of lead time, and when a 10 km buffer is applied, reduced improvement appeared as early as 25 minutes prior to an LSRs time of occurrence. These findings further highlight this potential limit of predictability within the ProbSevere model and introduce questions about why the automated guidance continues to miss LSR events even with a buffer applied, and what meteorological circumstances differentiate a missed event from a hit.

Despite the marked increase in the number of warning plume "hits" demonstrated across all lead times under the forecast-centric perspective, most improvement in the conditional POD of the warning plumes was seen across the longer lead times (Fig. 4.8a). At the time of the report, the cumulative fraction of LSRs warned by the automated PHI increased by about 2% with a 2

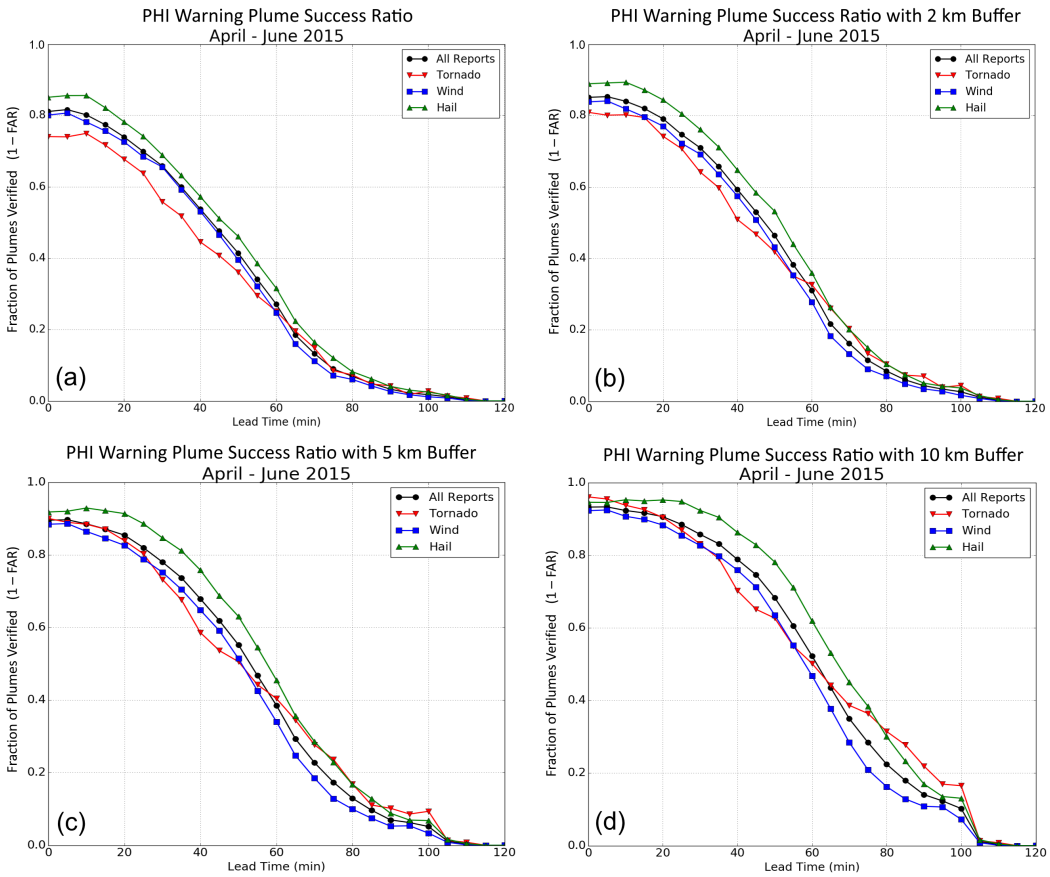

Figure 4.7: As in Fig. 4.3, but with a (a) 0 km, (b) 2 km, (c) 5 km, and (d) 10 km buffer applied to the automated ProbSevere objects and their derived PHI warning plumes.

km buffer, and by roughly 8% when a 10 km buffer was applied. However, these improvements generally increased with increasing lead time. Recall that this observation-centric perspective only considers the first warning plume to contain a given LSR when calculating lead time. As such, Fig. 4.8 shows the cumulative fraction of LSRs with a maximum lead time greater than or equal to a given lead time, as in Fig. 4.4. Therefore, this discrepancy in improvement seen under the forecast-centric and observation-centric perspectives likely indicates that although LSRs are being warned earlier and more consistently with a buffer applied, the automated guidance isnt necessarily catching more LSR events. In particular, the ProbSevere model is likely already predicting as many severe weather events at shorter lead times as it is currently able, regardless of buffer. These results also lend evidence to support the hypothesis that the geometric precision of the original ProbSevere objects may serve as another limit of predictability, such that the likelihood of providing extended lead time for a severe weather event is subject to a storms motion variability and other non-linear influences.

This buffer technique was also applied to SBWs, and the new cumulative POD was plotted as a function of lead time as shown in Fig. 4.8b. Similar to the PHI warning plumes, the cumulative POD of the SBWs showed little improvement with a buffer applied, with an increase of just 1% noted with a 10 km buffer at the time of the report. However, unlike the PHI plumes, SBWs did not exhibit any notable improvement at the longer lead times. These results can be largely explained by considering the potential predictability limitations affecting each warning method. Because the ProbSevere model produces storm objects for any storm that meets certain reflectivity and area thresholds, and because the experimental assumptions stated in section 4.1 provide PHI warnings for

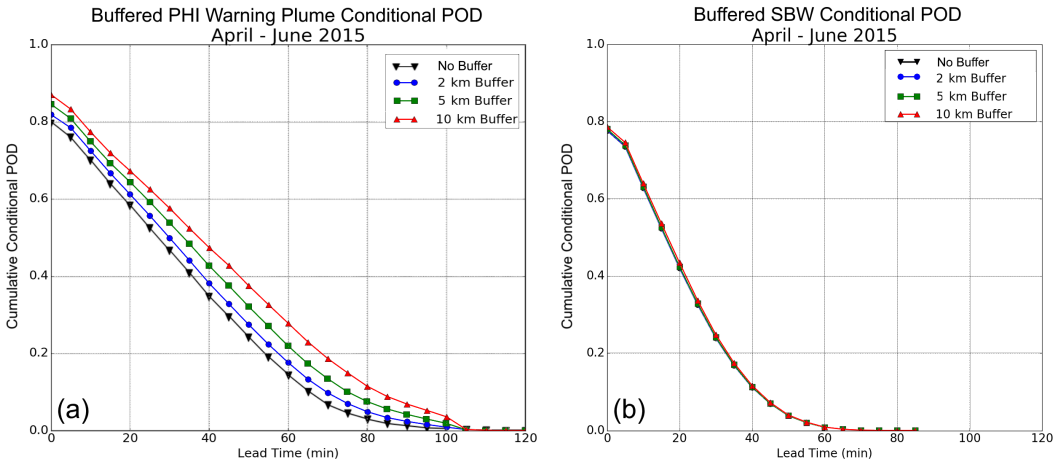

Figure 4.8: As in Fig. 4.4, but with a 0 km, 2 km, 5 km, and 10 km buffer applied to the automated ProbSevere objects and their derived PHI plumes (a) and SBWs (b).

any object associated with a storm that produced an LSR, most missed events are likely due to limitations of the storm tracking algorithms and variability in storm motion as discussed previously. However, the SBWs in this analysis were issued by human forecasters without prior knowledge of which storms would produce an LSR. Given the results shown in Fig. 4.8b, it is therefore likely that most events not warned by a SBW were missed due to limitations in scientific knowledge (Brooks 2004) rather than the geometries of the warning polygons.

New functionality added to the PHI prototype tool prior to the 2017 PHI experiment enabled forecasters to not only add a buffer around an automated object, but also to spatially shift the buffer such that the additional area is only added to one side of a storm. For instance, in the case of a classic tornadic supercell, forecasters could adjust the area of the object to better cover the hook echo and forward flank of the storm, where a tornado or damaging straight-line winds may occur. This could also be used to add warning area behind a storm

to ensure end-users remain in a warning until the threat has completely passed, or to indicate a storm may be changing direction or splitting, etc.

4.2.3 Area Statistics

Although this thesis did not attempt to calculate FAA performance metrics, a discussion of buffer verification would be incomplete without the inclusion of plume area statistics. In particular, the spatial areas swept out by unbuffered, automated PHI plumes conditionally verified by an LSR were calculated and compared to those of verified NWS-issued SBWs (Fig. 4.9). Overall, the automated plumes averaged an area of about 6,000 km^2, with a median area of 3,525 km^2. When stratified by report type (Fig. 4.9a), it can be seen that warning plumes conditionally verified by tornado LSRs covered the smallest area, with a median of 2,264 km^2. Plumes conditionally verified by hail LSRs were determined to have a median area of 2,663 km^2, and those verified by wind reports came in at 4,795 km^2. Most area distributions (represented by violin plots; Hintz and Nelson 1998) were skewed towards lower values, with peaks generally well below the median.

SBWs were smaller on average than the unbuffered PHI, with a mean area of about 2,200 km^2 and a median of 1,430 km^2 across all report types (Fig. 4.9b). Warnings verified by a tornado report (both severe thunderstorm and tornado SBWs) had the smallest median area of 951 km^2, followed by hail-verified warnings at 1,161 km^2. SBWs that verified with a damaging wind report had a median area of 1,835 km^2. These results suggest that SBWs provide a more precise warning area than the automated guidance. Perhaps this is to be expected as another limit of predictability, as it is reasonable to assume that

improvements in severe weather detection come at the expense of larger forecast areas and consequently FAA. However, another possible explanation may be related to the way PHI plumes are derived. As mentioned in section 4.1 and Karstens et al. (2015), the area swept out by a PHI warning plume is a function of a storms motion uncertainty, expected duration, and the geometric attributes of the plumes parent storm object. Furthermore, it was stated in section 4.2.1 that the automated PHI warning plumes are restricted to a maximum predicted duration of 120 minutes, or twice the length of a typical SBW duration. Therefore, it is hypothesized that PHI warning plumes may be spatially larger than SBWs simply because they are valid for longer periods of time.

To demonstrate this hypothesis, the polygonal areas provided by PHI warning plumes and SBWs were normalized by their duration as shown in Fig. 4.9c,d. As expected, verified PHI products were generally given much longer durations than SBWs, with mean values of about 77 and 45 minutes respectively across all report types. This difference did help to bring the PHI warning plume normalized areas closer to those of the operational SBWs, but SBWs continued to exhibit smaller normalized areas overall. For example, SBWs verified by any LSR had a median normalized area of 31 km^2/min while PHI warning plumes covered a median of 48 km^2/min. Even so, the normalized distributions represented by the violin plots in Fig. 4.9c,d demonstrate that the PHI warning plumes are comparable to SBWs when duration is taken into consideration. Thus, the tradeoff for increased lead time is an increase in implied warning area. Ultimately, it is likely a combination of predictability limitations, duration differences, paradigm differences, and the object identification methods used by ProbSevere that account for these discrepancies between SBWs and the PHI warning plumes. These quirks and limitations of the automated guidance

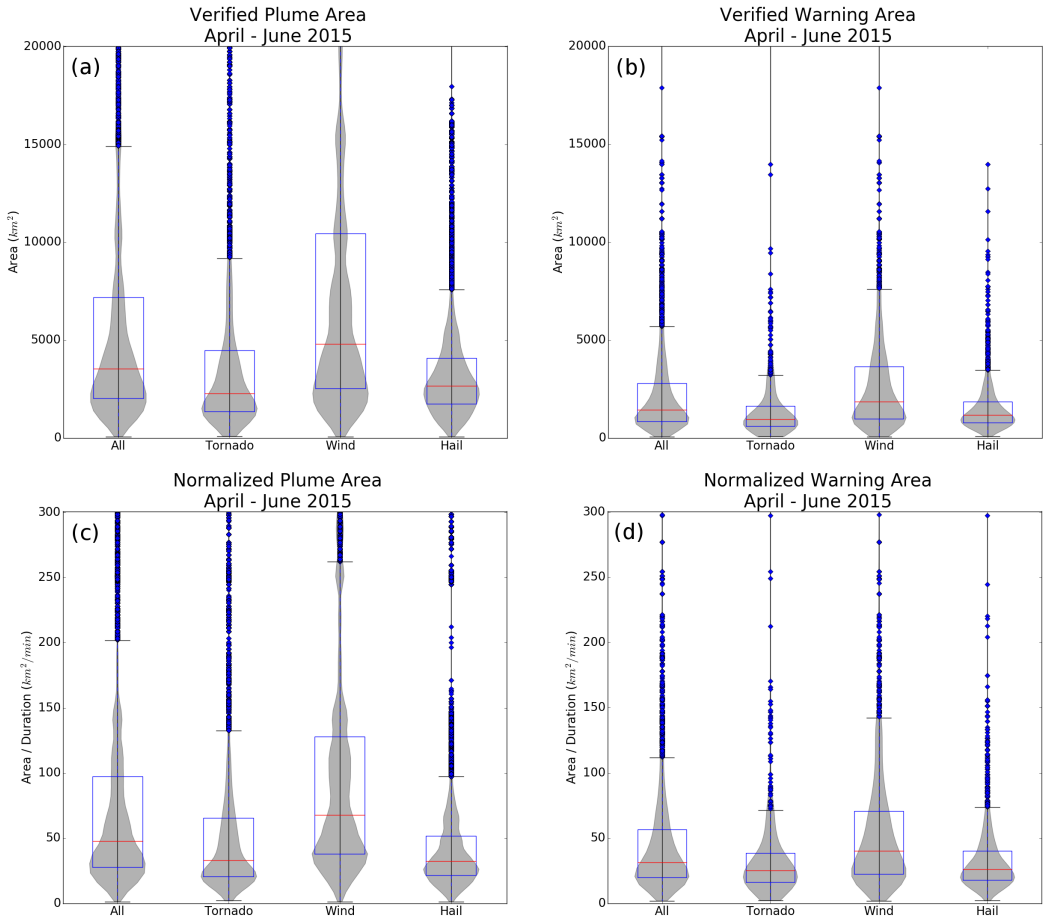

Figure 4.9: (a) PHI warning plume area distributions [violin plots (Hintz and Nelson 1998)] broken down by associated LSR type. (b) As in (a) but for SBWs. (c) Normalized PHI warning plume area distributions broken down by associated LSR type. Area is normalized by each plumes valid duration. (d) As in (c) but for SBWs.

leave forecasters with the challenge to not only correctly pick which storms will produce severe weather, but also to systematically adjust the associated warning plumes such that they cover the areas at risk while balancing over-warning considerations.

Chapter 5

Automated Probability Trend Predictions

As stated in section 2.2, the ProbSevere automated guidance produces the diagnostic probability that a storm will produce severe weather within the next 120 minutes. This value represents the total likelihood that severe weather will occur during the valid time frame, but does not provide information on how that probability may change with time. For example, a decaying thunderstorm may be more likely to be severe within the next 5 to 10 minutes than it is 40 minutes from the analysis time. Similarly, growing convection could have a low probability of producing severe weather in the near term, but be in an environment where it is likely that the storm will become severe later in its lifetime.

During the 2015, and 2016 PHI prototype experiments, participating NWS forecasters were provided the ProbSevere diagnostic probability that each storm object would be associated with severe weather. The forecasters were then asked to predict how these probabilities would change in time, and their predictions were used to populate gridded warning plume probabilities as described in section 2.3. Because ProbSevere provides no information on how these probabilities will change with time, forecasters were provided a limited first-guess prediction such that the initial probability of severe was set equal to the current ProbSevere diagnostic probability for each forecast. These probabilities were then reduced linearly with increasing forecast time, such that there would be a probability of 0% at the forecaster-specified end time of each plume (Fig. 5.1; Karstens

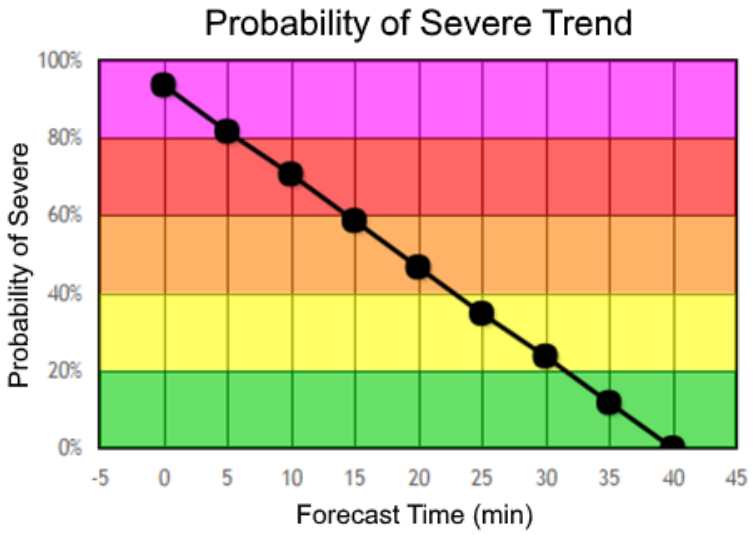

Figure 5.1: Default linear decrease from the current diagnostic ProbSevere probability to 0% by the end time of a warning plume.

et al. 2015). In addition, functionality was included within the PHI prototype interface that allowed forecasters to adjust these probability trends by manually changing the probability at each time step, or by setting an initial value and a final value which could then be used to automatically fit a positive exponential decay, negative exponential growth, or bell curve function to those points to account for subjectively-derived dynamic evolution similar to the scenarios described previously (i.e., decaying thunderstorm).

Karstens et al. (2015) found using a grid-based method, not replicated in this thesis, that forecasters participating in the 2014 PHI prototype experiment showed some reliability when issuing lower probability forecasts, but their predictions generally tended to over-forecast in the mid- to high-probability spectrum (Fig. 5.2). In fact, the participating forecasters' average reliability fell near climatology, and well below the skill-no-skill line during the week-long

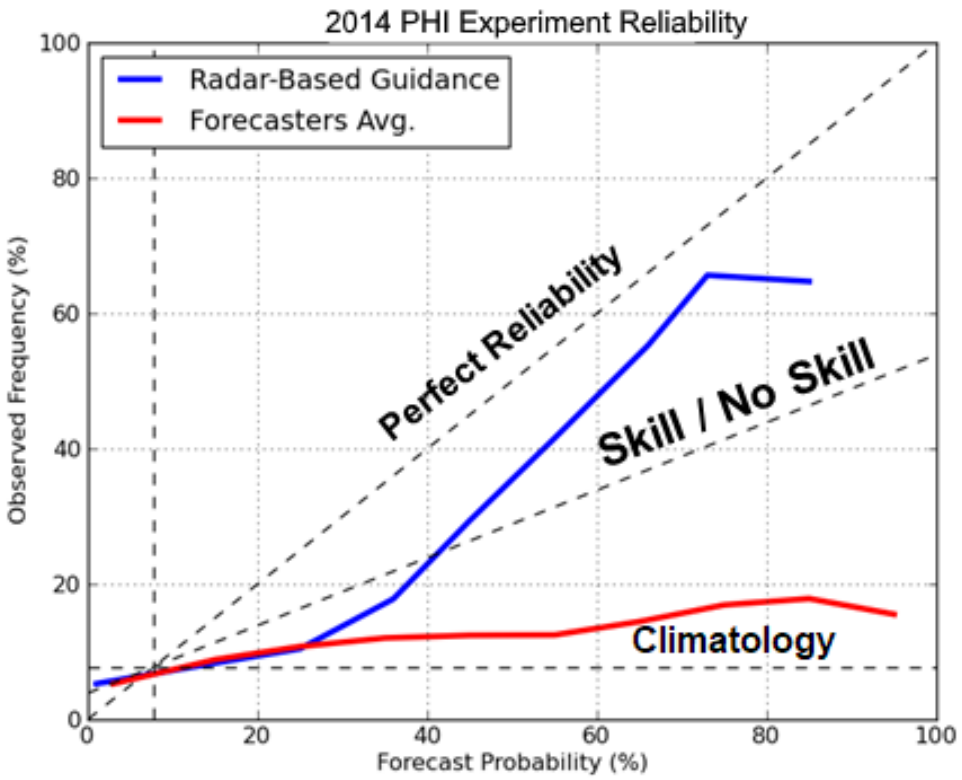

Figure 5.2: Reliability diagram displaying the mean probabilistic forecasts of all forecasters participating in the 2014 PHI prototype experiment, verified using MESH values ≥ 1 in. For comparison, probabilistic forecasts were generated using a linear decay rate for the recommended probability value. Modified from Karstens et al. (2015), their Fig. 15a.

experiments. In contrast, the diagnostic predictions from an early prototype probability model similar to ProbSevere exhibited reliability that generally fell between the skill-no-skill line and perfect reliability. The model also tended to over-forecast at higher probabilities, but at a rate much less than the forecasters. Note that the skill-no-skill line falls halfway between climatology and perfect reliability.

Feedback from this and later PHI prototype experiments suggest that NWS forecasters lack direct experience with how a storm's probability of producing severe weather may evolve in different convective situations. To address this limitation, the second goal of this thesis is to use machine learning techniques to produce a first-guess model capable of predicting how a storm's probability of being severe will change with time. I hypothesize that such a model could facilitate an improved first-guess probability trend based on particular meteorological situations rather than a default linear decay, and potentially improve forecaster interdependence with the automated guidance.

5.1 Training Data

Any model or process that modifies data associated with a ProbSevere storm object as part of the experimental PHI prototype tool is required to complete its task in less time than the ProbSevere update interval (~2 minutes). To meet this requirement, it was therefore necessary to only use data that is readily available at runtime for training, and only variables provided by or derived from the BTRT-corrected ProbSevere model output (see section 2.2) were considered in this thesis. In particular, the time-of-analysis MUCAPE, EBS, and MESH

Description	Data Fields
RAP Fields	MUCAPE, EBS, Δ_{t-1}MUCAPE, Δ_{t-1}EBS
Storm Attributes	MESH, $\Delta\epsilon_{tot}$, Δice, Δ_{t-1}MESH, $\Delta_{t-1}\Delta\epsilon_{tot}$, $\Delta_{t-1}\Delta$ice
Geometric Attributes	Area, Perimeter, Δ_{t-1}Area, Δ_{t-1}Perimeter
Metadata	Direction, Duration, Velocity South, Velocity East, Δ_{t-1}Direction, Δ_{t-1}Duration, Δ_{t-1}Velocity South, Δ_{t-1}Velocity East
ProbSevere Probability	Current Prob, Max Prob, Prob Variance, Δ_{t-1}Prob, Δ_{t-2}Prob

Table 5.1: Inputs used to train each machine learning model to predict how the probability of a storm being severe will change with time. Δ_{t-1} indicates the change in a variable between the current and previous time step (~2 min).

were extracted from the radar-based ProbSevere storm objects, and the glaciation rate (Δice) and growth rate ($\Delta\epsilon_{tot}$) were gathered from the satellite-based objects to provide a representation of the atmospheric environment of each storm. In addition, geometric attributes, such as an object's polygonal area and perimeter, were derived from the radar-based storm objects to represent physical changes in a storm. Metadata for each object, including a first-guess storm duration prediction provided by McGovern et al. (2017, 2018b) and the storm's motion vector were also included in the training dataset. Finally, the current ProbSevere diagnostic probability was provided as input for training to act as an initial boundary condition for the machine learning model.

Short temporal trends for each of these variables were created by calculating the change in each value between the previous and current ProbSevere model

output (~2 min). If a storm object had only been in existence for a single time step at the time of analysis, then the temporal change was set to 0 for each variable. In addition, the change in the ProbSevere diagnostic probability for a given object was calculated for two time steps prior to the current time (~4 min). The maximum diagnostic probability and variance was then calculated for the entire lifespan of each storm up to the time of analysis. This resulted in 29 total variables available to use as inputs for training a machine learning model. A full list of the training input variables is provided in Table 5.1.

5.2 Machine Learning Algorithms

A combination of linear and ensemble regression learning algorithms were trained and compared in this thesis, including a random forest regressor, AdaBoost regressor, gradient boosting regressor, and an elastic net. In addition, an isotonic regressor was applied to the output of each of those models in an attempt to correct for model bias. These machine learning algorithms and techniques are briefly described in the following subsections.

5.2.1 Random Forests

A random forest (Breiman 2001) is a collection of classification or regression trees (Breiman 1984) in which each tree is trained on data that is bootstrap resampled with replacement from the training dataset, and a small random subset of the total number of training variables are evaluated for splitting at each node within each tree. This random sampling method forces tree nodes to split along the best variable in the subset instead of the best overall variable, meaning that some trees in the forest are grown from suboptimal features within

a dataset. As a result, random forests grown in this way exhibit greater tree diversity, which trades a higher model bias for lower variance compared to other tree-building techniques (Géron 2017). Once a sufficient number of trees have been grown from the input data, the final prediction from the forest is output as the mean of the predicted values from all individual trees. By averaging the results from all trees in this manner, random forests are able to produce a smoother range of predicted values with lower variance than a single decision tree (Strobl et al. 2008).

5.2.2 AdaBoost

Géron (2017) defines boosting as any ensemble method that can combine several weak learners into a strong learner. Specifically, most boosting techniques train each iteration of a model sequentially, with each iteration trying to correct the errors of the previous iteration (Freund and Schapire 1997; Drucker 1997). The Adaptive Boosting technique (AdaBoost; Freund et al. 1996; Freund and Schapire 1997; Friedman 2001, 2002) follows this approach by first building and training a base regression model (typically a decision tree) to make predictions on an input training dataset. The error in the prediction of each training instance is then calculated, and relative weights are assigned such that training samples with larger error are given a larger weight. AdaBoost then trains another regression model on the same input data, but with the weights applied such that training instances with large errors in the previous iteration contribute more to the current predictor's cost function. The weights are then updated based on the performance of the new model. This method effectively forces each iteration of AdaBoost to focus more on correctly predicting training samples that had a large error in the previous iteration.

Once a specified number of predictors have been trained, a new set of weights is applied to each predictor, such that those with large overall error on the weighted training dataset are weighted less than those with low overall error. Final predictions from AdaBoost are then computed similarly to those of random forests, except instead of treating each predictor equally and taking an ensemble average, AdaBoost values the predictions of each predictor using the weights assigned in the final step of the algorithm. This means that predictors that had lower overall error on the training set contribute the most when making the final predictions.

5.2.3 Gradient Boosting

As the name suggests, gradient boosting (Breiman 1997; Friedman 2001, 2002) is another type of boosting technique. Similar to AdaBoost, gradient boosting sequentially adds predictors to an ensemble, with each new member trained to correct the errors of the previous iteration. However, where AdaBoost applies weights to the training data in an attempt to reduce the errors of each new iteration, gradient boosting trains each new predictor on the residual errors of the previous iteration (Géron 2017). For example, consider a regression decision tree trained on some arbitrary dataset. The training set residual error for that tree is the difference between the true value of each sample and the predicted value for that sample. Using the gradient boosting technique, a new decision tree is then fit to the training dataset such that the residual errors of the previous tree are the expected output for each training sample. The training set residual errors for the new predictor are then calculated and the process is repeated. Once a specified number of predictors have been trained, the final prediction of a gradient boosted ensemble is then simply the sum of the predictions of all

predictors in the ensemble. One disadvantage of gradient boosting is that is generally takes longer to train than the other methods discussed in this thesis because the learning processes cannot be parallelized. That is, each tree must be trained sequentially, and only one tree can be grown at a time. However, several studies (e.g. McGovern et al. 2017; Gagne II et al. 2009) have shown that gradient boosting may produce more robust models that are able to better generalize to noisy input data than other learning techniques.

5.2.4 Elastic Nets

In contrast to the previous three learning techniques, an elastic net (Zou and Hastie 2005) is a type of linear regression model that constrains the weights of each input variable to regularize the output of the model by reducing the number of polynomial degrees (Géron 2017). This regularization forces the model to not only fit the data, but to keep the weights of each input feature as small as possible to avoid overfitting. In particular, elastic nets combine the regularization techniques of Ridge regression (Hoerl and Kennard 1970) and least absolute shrinkage and selection operator (Lasso; Tibshirani 1996) regression models, such that the cost function for an elastic net can be represented as

$$J(\theta) = MSE(\theta) + r\alpha \sum_{i=1}^{n} |\theta_i| + \frac{1-r}{2}\alpha \sum_{i=1}^{n} \theta_i^2. \tag{5.1}$$

Here, r is the mix ratio, where $r = 0$ gives a cost function equivalent to that of a Ridge regression model and $r = 1$ is equivalent to lasso regression. The hyperparameter α determines how much regularization to apply, and θ represents the weights of each input variable. The primary practical difference between ridge regression and Lasso regression regularization techniques is that Lasso regularization tends to completely eliminate the weights of the least important features

by setting them to 0 (Géron 2017). By incorporating a blend of both regularization techniques, elastic nets are able to outperform both Lasso and Ridge regression individually, particularly when the number of training variables is much larger than the number of training samples (Zou and Hastie 2005).

5.2.5 Isotonic Regression

In this research, each of the previous machine learning techniques produces a probability value representing the likelihood of severe weather at a given forecast time. However, anecdotal test cases found that these models can be prone to over or under forecasting the probability of a severe storm in various circumstances, resulting a reliability graph that is systematically offset from the one-to-one line (Wilks 2011). To account for this, an isotonic regressor was used to attempt to correct any model bias in this analysis.

Isotonic regression (Dykstra and Robertson 1982) is a form of statistical inference (Barlow 1972) that is restricted by the order of the data. Specifically, an isotonic regressor finds a non-decreasing approximation of a function while minimizing the mean squared error of the training data[1]. This type of model does not assume linearity or any other specific form for the target function, except that each point of the function must be greater than or equal to the previous point (non-decreasing). In practice, isotonic regression is used as a calibration technique which trains on the probabilistic output from another model in order to adjust those probabilities to more closely follow the one-to-one line on a reliability graph (Zadrozny and Elkan, 2002). Isotonic regression was applied to the output of each of the previously discussed models as described in section 5.3.

[1] http://scikit-learn.org/stable/auto_examples/plot_isotonic_regression.html

5.3 Machine Learning Methods and Procedures

As stated in Chapters 3 and 4, there were a total of 1,520,330 unique BTRT-corrected ProbSevere storm objects during the spring and early summer months (April - June) of 2015 that were available to train and test machine learning models on for this thesis. However, because the ProbSevere model produces storm objects for any storm that meets minimum intensity and area thresholds (section 2.2), a disproportionate number of storm objects in the dataset had very low ($< 10\%$) probabilities of producing severe weather at any point in their lifetimes (Fig. 5.3). Because of this, it was necessary to first apply an undersampling technique to balance the training dataset.

Undesampling was performed by first calculating the maximum diagnostic probability of severe that ProbSevere predicted during the entire lifetime of each storm object. That is, the diagnostic probability of severe for every storm object at every analysis time in that storm's lifetime was compared, and the maximum value was taken as the maximum probability for that storm. All storm objects were then stratified by their maximum probability values into 10% bins. In order to produce a well-calibrated model, the training input dataset required an approximately equal representation of storms in each maximum probability bin. Therefore, the undersampled dataset was created by randomly sampling without replacement from each bin, until the number of samples drawn from each bin equaled the number of samples in the smallest bin (Fig. 5.3). The final training dataset used for training the base models was then made up of each individual analysis time of every automated ProbSevere storm object in the undersampled dataset.

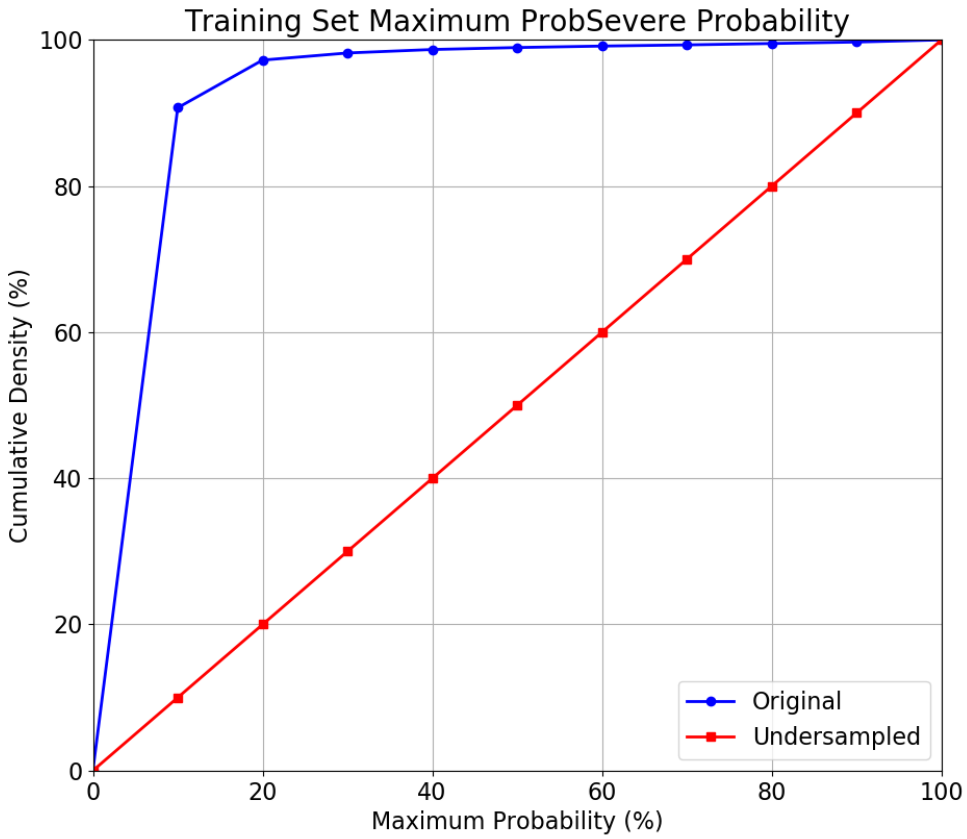

Figure 5.3: Cumulative density functions of the maximum ProbSevere diagnostic probability prediction of the original dataset (blue) and the undersampled dataset (red). The data was binned in 10% increments, such that the cumulative density reported at a given probability is the percent of ProbSevere objects with a maximum ProbSevere probability prediction $\leq$ that probability.

Recall that each bin is based on the maximum probability of severe over a storm's entire lifetime. However, this does not take into account any of the individual probabilities predicted at a given analysis time for each storm. Anecdotally, severe thunderstorms tend to have lower ProbSevere probabilities early in their lifecycle and then gradually or rapidly ramp up to higher severe probabilities as the storms mature. As such, even storms with a higher maximum probability of severe have lower diagnostic probabilities for part of their lifecycle. Because every individual analysis time of a ProbSevere storm object was included separately in the undersampled training data, this means that there were still more individual samples with lower diagnostic ProbSevere probabilities. Therefore, all models trained on the undersampled dataset were still exposed to imbalanced probabilities but at a greatly reduced rate than in the original data.

K-fold cross validation (Kohavi et al. 1995) was used to train and tune hyperparameters for a random forest, AdaBoost regressor, gradient boosting regressor, and an elastic net. ProbSevere objects from 9 April 2015 to 1 May 2015 were chosen as the training and validation dataset and partitioned into five folds composed of five sequential days of data. At least one day of data was withheld between each fold to preserve sample independence. Each model and hyperparameter choice was then trained on data from four of the folds and validated on the remaining fold. This process was repeated five times for each model (5-fold cross validation), such that each fold was used for validation exactly once per model. Finally, the mean absolute error (MAE) was calculated for each validation fold, and the average MAE of all five validation folds was reported as the validation performance for that model and hyperparameter selection. Confidence intervals were also computed for each model using the

bootstrapping percentile method (Efron and Tibshirani 1986). This technique randomly chooses samples with replacement from each validation fold and then calculates the absolute error of each prediction in the bootstrapped data. The 95% confidence interval is then reported as the upper and lower 2.5% points of the distribution, where 95% of the absolute error of the bootstrapped predictions falls between those values (Efron and Tibshirani 1986).

The best hyperparameters for each model were determined by identifying the hyperparameter selection that resulted in the lowest validation MAE and smallest confidence interval. An isotonic regressor was then trained on the predictions of each model as described in section 5.2.5. ProbSevere objects from 1 May 2015 to 1 June 2015 were used to perform this model calibration, again using 5-fold cross validation with five sequential days of data in each fold. As before, more than 90% of the ProbSevere objects in the calibration dataset had maximum probabilities of less than 10%, so undersampling was again performed to avoid under-predicting the more severe storms. Finally, the best models trained by each of the four learning techniques with and without the isotonic regression calibration were tested on ProbSevere objects from 1 June 2015 to 1 July 2015, and the MAE and confidence intervals were reported. Early tests indicated that the models generally did not perform well on ProbSevere objects with a diagnostic probability $\leq 15\%$, likely due to the undersampling methods used for training. Furthermore, results and anecdotal observations from the 2015 and 2016 PHI prototype experiments indicate that forecasters generally don't directly interact with ProbSevere storm objects with a diagnostic probability $\leq 15\%$. Therefore only objects with a diagnostic probability $> 15\%$ at a given time step were included in the validation and test sets to better represent how the models may perform in an operational environment.

5.4 Model Performance

Validation scores and confidence intervals for the best hyperparameter selections for each model and isotonic regression are provided in Tables 5.2 and 5.3. In general, each model exhibited similar mean performance on the validation set data, with MAEs ranging from 17.7% with the gradient boosting regressor to 19.3% for the random forest model. Validation set error was reduced for all models when only considering forecast times $\leq$ 10 min., with the gradient boosting regressor once again providing the lowest MAE at 8.19%. Similarly, MAEs increased when only considering forecast times $>$ 10 min. Interestingly, the models calibrated by the isotonic regression did not perform notably better than the uncalibrated models. In fact, the validation set MAE actually increased slightly for each model, both at short and long forecast times. One possible explanation for this is that the initial models were already relatively reliable and did not need additional calibration. In this case, the isotonic regression may have overfit the predictions of the base models, resulting in increased error.

While each model and isotonic regression generally had a validation set MAE around or less than 20%, the 95% confidence intervals for each model varied more noticeably. For example, the gradient boosting regressor exhibited a mean validation set confidence interval of 0.007% - 68.8%, while the AdaBoost regressor had a confidence interval of 0.122% - 56.2%. Overall, each model exhibited validation set confidence intervals of approximately 0% - 65%, though the upper bound reduced to about 40% when only considering short forecast times. These results indicate that although the models are able to make forecasts with relatively low error on average, they still occasionally produce probability

predictions with large absolute error compared to the ProbSevere diagnostic output. One possible explanation for this can be traced back to how ProbSevere produces its diagnostic probabilities. Recall from section 2.2 that the ProbSevere model is a naive Bayesian classifier which primarily relies on the most recent satellite and radar observations to produce diagnostic probability assessments. This means that the ProbSevere probabilities are sensitive to rapid changes in a storm or its surrounding environment, and as such, the ProbSevere diagnostic probabilities often exhibit large changes over short time periods. In contrast, the machine learning models in this thesis are trained on both current observations and observational trends from the past one to two time steps. By including temporal trends for each variable as inputs into the models, each machine learning algorithm is trained in a way that essentially smooths over changes in the diagnostic ProbSevere probabilities as shown in Fig. 5.4. Additionally, all machine learning algorithms utilized in this thesis except for the Elastic Net are ensemble regressors which inherently produce a smoother range of probability predictions than an individual decision tree or statistical model (Strobl et al. 2008). These smoothing effects are likely partially responsible for the large confidence intervals of the machine learning models.

Because each model exhibited similar performance such that no one model was significantly better than the others, an ensemble approach was chosen to make the final probability predictions. In this method, each model separately predicts the probability that a storm will be severe at each forecast time, and these predictions are then averaged to produce the final ensemble prediction. Numerous studies (e.g. Hamill et al. 2000, 2003; Lewis 2005) have demonstrated the benefits of ensemble approaches over the use of a single model for diagnostic and probabilistic forecasts. For example, Whitaker et al. (2006) noted that

	Random Forest	**AdaBoost**	**Gradient Boosting**	**Elastic Net**
Base				
All Times	19.3	18.6	17.7	18.9
Short Times	9.56	10.2	8.19	14.8
Long Times	22.4	23.6	20.4	22.5
Isotonic				
All Times	20.7	20.9	20.5	21.4
Short Times	10.6	10.2	11.8	14.9
Long Times	23.9	24.3	23.2	23.5

Table 5.2: Validation set MAE for each base regressor and with isotonic regression applied at all forecast times, short forecast times ($\leq$ 10 min.), and long forecast times ($>$ 10 min.)

	Random Forest	**AdaBoost**	**Gradient Boosting**	**Elastic Net**
Base				
All Times	.417 - 61.6	.122 - 56.2	.007 - 68.8	1.08 - 57.6
Short Times	.044 - 41.3	0.00 - 39.9	0.00 - 43.1	.662 - 42.1
Long Times	1.24 - 63.4	1.72 - 57.8	.658 - 70.9	1.36 - 59.7
Isotonic				
All Times	.544 - 61.4	.619 - 58.4	1.07 - 59.4	1.10 - 59.1
Short Times	.221 - 41.8	.186 - 40.8	.598 - 41.0	.542 - 43.3
Long Times	1.54 - 63.5	1.60 - 60.0	1.32 -64.4	1.51 - 61.2

Table 5.3: Bootstrapped validation set confidence intervals for each base regressor and with isotonic regression applied at all forecast times, short forecast times ($\leq$ 10 min.), and long forecast times ($>$ 10 min.)

	Random Forest	**AdaBoost**	**Gradient Boosting**	**Elastic Net**	**Ensemble**
Base					
All Times	20.7	21.5	18.9	21.9	20.4
Short Times	10.6	11.1	9.24	15.6	11.3
Long Times	24.0	24.8	21.9	24.0	23.3
Isotonic					
All Times	21.0	20.9	20.7	22.0	
Short Times	10.8	10.7	11.9	15.5	
Long Times	24.3	24.2	23.5	24.0	

Table 5.4: As in Table 5.2, but for test set MAE. The ensemble MAE is calculated by averaging the predictions of all eight models at each forecast time.

	Random Forest	**AdaBoost**	**Gradient Boosting**	**Elastic Net**	**Ensemble**
Base					
All Times	.379 - 62.2	.111 - 56.3	.003 - 69.2	1.19 - 56.7	.930 - 58.6
Short Times	.015 - 42.2	0.00 - 40.2	0.00 - 44.0	.662 - 43.0	.398 - 40.7
Long Times	1.53 - 64.5	1.96 - 57.9	.794 - 71.6	1.67 - 58.7	1.69 - 60.6
Isotonic					
All Times	.413 - 62.2	.579 - 57.3	1.09 - 62.4	1.08 - 59.4	
Short Times	.064 - 42.1	.300 - 41.2	.825 - 40.7	.589 - 43.7	
Long Times	1.70 - 64.4	1.68 - 59.0	1.37 - 64.6	1.67 - 61.3	

Table 5.5: As in Table 5.3, but for test set MAE. Ensemble confidence intervals are calculated by averaging the predictions of each model at each forecast time.

ensembles of forecasts with a single model suffer from a deficiency in spread, partly because they do not represent the error in the forecast model itself. Their observation is of particular importance to this research, as random forest, AdaBoost, and gradient boosting regressors are all ensemble learning algorithms (see section 5.2); therefore, the predictions from any one of these models are likely to be biased by the errors of that model. However, by taking an ensemble of all four base regressors and their isontonic calibrations, it is possible to incorporate the strengths of each model to better balance out their inherent biases.

The test set MAE and confidence intervals of each calibrated and uncalibrated model, as well as for the ensemble as a whole, are provided in Tables 5.4 and 5.5. Overall, each individual model exhibited similar test set performance to that reported for the validation set, with the gradient boosting regressor again reporting slightly lower MAE than the other individual models. Averaging the predictions from each model resulted in a test set MAE of 20.4% for the ensemble, with a 95% confidence interval of 0.930% - 58.6%. Note that the ensemble test set MAE outperformed all but two of the eight individual members. Similarly, the ensemble 95% confidence interval was smaller than all but that of the AdaBoost regressor and its isotonic calibration. These results further support the use of an ensemble rather than an individual model, as the ensemble was able to achieve MAE near to that of the gradient boosting regressor (low MAE, large confidence interval), with a confidence interval similar to what was achieved by the AdaBoost regressor (high MAE, small confidence interval).

A series of case studies were also performed to further evaluate the ensemble. In these studies, the ensemble was provided with a single ProbSevere storm object at a single analysis time and asked to predict the probability that the

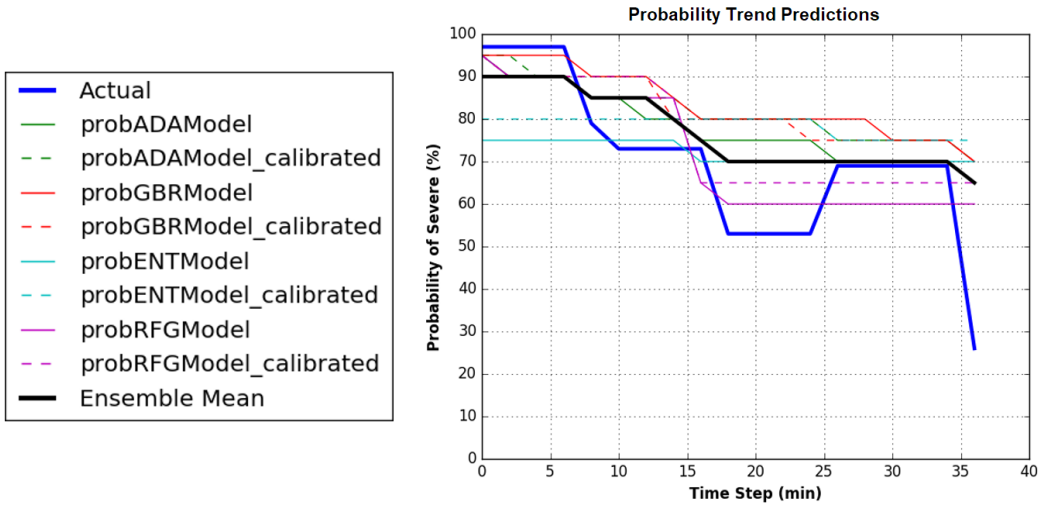

Figure 5.4: Probability predictions for a single ProbSevere storm object with a predicted duration of 37 minutes at a single analysis time. The ensemble average for each forecast time is shown in a thick black, and the actual diagnostic ProbSevere probability for the corresponding time is shown in a thick blue.

storm would be severe in 5 minute forecast intervals between the current analysis time and the predicted end time of the storm (McGovern et al. 2017, 2018b). The predictions from each individual model and the ensemble mean were then plotted and compared to the true diagnostic ProbSevere probabilities at the corresponding times (Fig. 5.4). These analyses anecdotally revealed that the ensemble is generally successful at predicting the overall temporal trend of the ProbSevere diagnostic probabilities, but often misses large changes in probabilities over a short time span. In fact, the ensemble tends to essentially smooth out the diagnostic ProbSevere probabilities as discussed previously, but this is not necessarily undesired behavior. Recall that the purpose of this research is to provide NWS forecasters with a first-guess probability trend to aid in the issuance of probabilistic severe weather warnings. Therefore, it may not be

necessary to capture every short-term change in a storm's severe probability as long as the general trend is well forecasted. These results indicate the need to test the ensemble predictions in an operational environment where human forecasters can subjectively evaluate the accuracy and value of the first-guess predictions. To acquire these metrics, the ensemble was implemented in the PHI prototype tool during the 2017 PHI prototype experiment, and the results from the participating NWS forecasters are provided in Chapter 6.

Chapter 6

Results from the 2017 Hazardous Weather Testbed

The underlying theme of this thesis is to improve forecaster-guidance interdependence in a new probabilistic warning paradigm. This ultimate goal was pursued by providing verification metrics for the automated ProbSevere guidance (Chapter 4) and designing a new model to provide forecasters with a first-guess probability trend to use as a starting point when issuing probabilistic forecasts (Chapter 5). Results from each of these tasks were determined experimentally using assumptions to simulate how forecasters may use the PHI warning system operationally. However, human input is needed in order to truly gauge the success of this research; therefore NWS forecasters participating in the 2017 PHI prototype experiment at the NOAA HWT were provided with the verification results from Chapter 4 and the new probability trend models from Chapter 5. This chapter focuses on how the forecasters utilized this new information and functionality to issue experimental PHI warning plumes while balancing the time constraints of operations during active severe weather events.

6.1 Testbed Design

The 2017 PHI prototype experiment consisted of three weeks of experimentation, in which nine NWS forecasters were given the chance to issue experimental PHI warnings on a series of three archived case studies and displaced real-time events as described in Karstens et al. (2018). However, because the displaced real-time events varied significantly from week to week, only results from the archived case studies are presented herein. Each case study lasted for approximately two hours and featured high-end severe weather events in the central and southern Plains, including isolated supercells and supercell clusters that produced long-track tornadoes and significant hail. During the course of the experiment, each forecaster was exposed to a total of 41 LSRs, and 2,428 warning plumes were issued for 186 automated ProbSevere storm objects across the three weeks. Note that this is a significantly smaller sample size than what was used in the rest of this thesis, and this should be taken into consideration when interpreting the results presented herein.

6.2 Forecaster and Model Predictions

6.2.1 Plume Verification

Chapter 4 focused on the verification of PHI warning plumes conditioned by the assumption that a hypothetical forecaster is able to always issue warnings that contain or are very near to an LSR at the earliest possible lead time (section 4.1). However, the real world is rarely perfect, and the experimental assumptions presented previously are unable to capture the human element that a real forecaster provides in an operational setting. Therefore, this subsection

offers a brief examination of how NWS forecaster-issued warning plumes verified when utilizing the automated ProbSevere guidance during the 2017 PHI prototype experiment.

To better identify how the forecasters utilized the automated guidance, each PHI warning plume was triplicated and post-processed into one of three categories: "original", "no buffer", and "automation only". As the names suggest, warnings in the "original" category were analyzed exactly as the forecasters issued them, including any changes to the automated guidance such as spatial buffers or changes to the motion vector. Warnings in the "no buffer" category were stripped of any spatial buffer the forecasters may have applied, but all other changes to the automation were left in place. Finally, warnings in the automation only category had all forecaster modifications removed such that the warning plume geometry and attributes were exactly what the automated guidance produced. For comparison purposes, the forecaster-specified warning durations were left the same across all three categories. Most warning plumes had durations of 45 minutes or less, and no warnings were issued for longer than 60 minutes, regardless of the longer-recommended duration times provided by the automation.

Interestingly, warning plume performance showed little difference between the three categories overall. Under the forecast-centric perspective, about 75% of the "original" forecaster-issued plumes contained an LSR at the time of occurrence, while about 69% and 70% of the "no buffer" and "automation only" plumes contained an LSR at the same lead time, respectively (FAR of 25%, 31%, and 30% respectively; Fig. 6.1a). Warning plumes governed only by the automated guidance had more hits than the forecaster-issued warnings between 20 and 30 minutes of lead time, but the forecaster-issued warnings generally

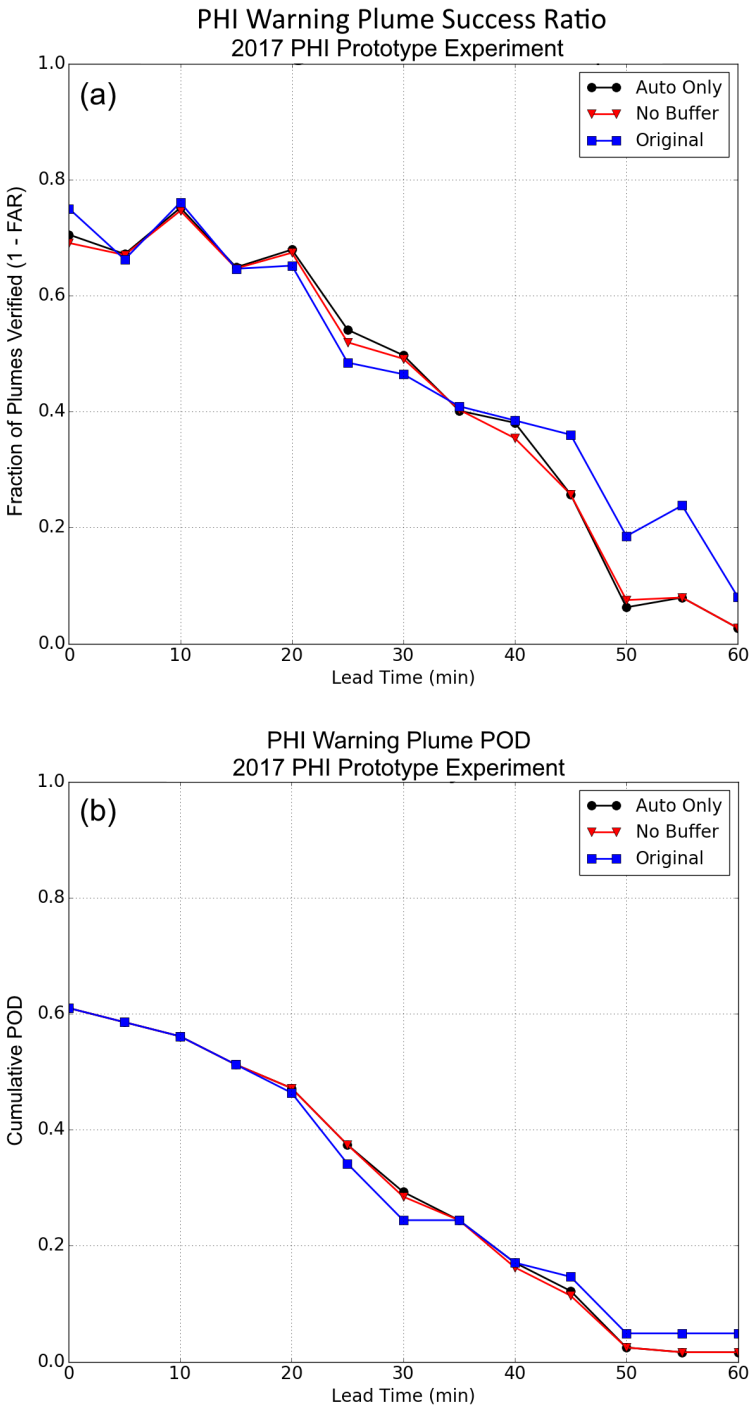

Figure 6.1: (a) As in Fig. 4.3 and (b) as in Fig. 4.4 but both for PHI warning plumes issued by forecasters during the 2017 PHI prototype experiment.

performed better more than 35 minutes prior to an LSR. Similar results were noted under the observation-centric perspective, with warning plume POD essentially identical across all three categories within 15 minutes of an LSR (POD of 0.6 at the time of an LSR). Again, the "automation only" plumes had a slightly higher POD between 20 and 35 minutes of lead time, while the "original" forecaster-issued warning plumes performed slightly better more than 40 minutes prior to an LSR.

Note that both the success ratio and POD in Figs. 6.1a and 6.1b were lower overall during the 2017 PHI prototype experiment than during the three-month period analyzed earlier in Chapter 4. This may be partially due to the limited sample size from the PHI prototype experiment, as well as differences in the experimental conditions and methods applied to the two datasets. Regardless, these results are encouraging, as they show that forecasters not only utilized the automated guidance during the experiment, but they also added value to the forecasts, particularly at the longer lead times. Ultimately, this appears to lend support toward improving forecaster interdependence with the automated guidance, and demonstrates that forecasters can use the automated guidance as a first-guess and identify situations requiring adjustments that can improve forecast quality.

6.2.2 Probability Model Performance

Forecasters' use of the new first-guess probability trend predictions was also analyzed as part of the 2017 PHI prototype experiment. For this analysis, participants were asked to issue PHI warning plumes as described in section 2.3, but the mean ensemble probability trend predictions were provided in place of the default linear decrease (Fig. 6.2). Forecasters retained the ability to modify

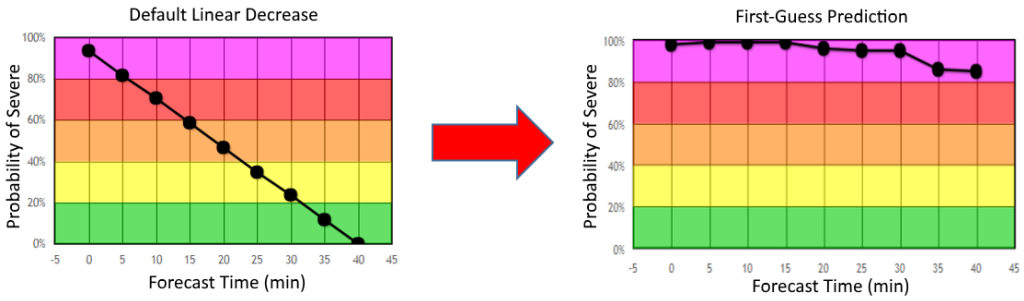

Figure 6.2: Example showing the difference between the default linear decrease probability trend provided to forecasters in 2016 (left), and the first-guess probability trend predictions provided in 2017 (right).

these probabilities as needed, or they could apply one of the aforementioned smoothing functions to the predictions as well. It was noted that the random forest ensemble member required more time to return probability trend predictions than the update interval of ProbSevere (about 2 minutes). Therefore, the random forest was removed from the ensemble in order to provide probability trend predictions in real time during the experiment.

The primary objective of this analysis was to determine how much the NWS forecasters changed the probability trend predictions produced by the first-guess guidance, and to compare those changes with the diagnostic ProbSevere probabilities. Karstens et al. (2018) showed that there was a significant change in forecaster use of the provided probabilistic information between the 2016 and 2017 experiments, noting that forecasters in the 2017 PHI prototype experiment used the ProbSevere diagnostic probability as the initial value in their trend predictions approximately six times more than participants in the 2016 experiment (Fig. 6.3). Additionally, forecasters generally did not make significant changes to the first-guess probability trend predictions in 2017, and most adjustments to the predictions were to increase the probability of severe at a given forecast

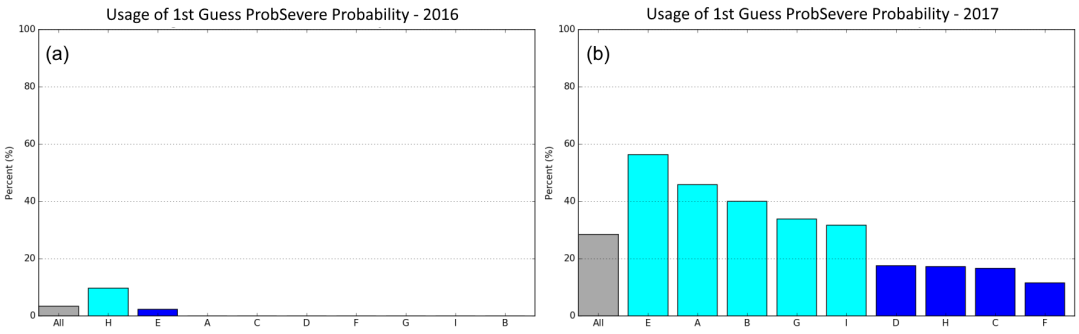

Figure 6.3: Percent usage of the first-guess diagnostic probability from ProbSevere objects by forecasters in (a) 2016 and (b) 2017, where each bar represents an idividual forecaster, light-blue shading represents above average usage, and dark-blue shading represents below average usage for that year. Modified from Karstens et al. (2018), their Fig. 12.

time (Fig. 6.4). For comparison, the median forecaster-issued and automated first-guess probability predictions for each forecast time are shown in Fig. 6.4a, where the shaded regions represent the 25th and 75th percentiles of each data distribution. These percentiles were calculated by randomly sampling forecaster and guidance predictions with replacement to create a bootstrapped dataset of 1000 samples each. The original default linear decrease was also calculated for each forecast and plotted as the others.

Overall, the probability predictions provided by the first-guess ensemble did not vary much with increasing forecast time, with a median probability of severe of about 90% at the time of analysis, 77% at a forecast time of 20 minutes, and 81% at a forecast time of 50 minutes. In contrast, median forecaster-issued probability predictions steadily decreased from about 95% at warning issuance to 88% at a forecast time of 20 minutes, before dropping sharply beyond a forecast time of 45 minutes. This resulted in a median difference between

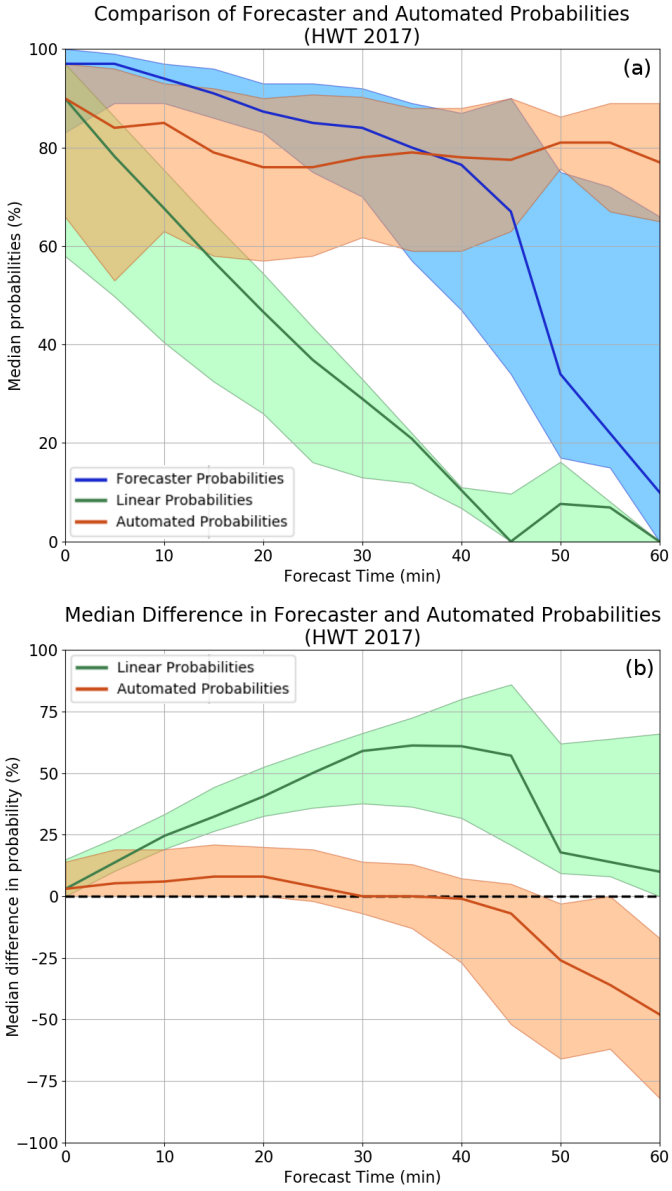

Figure 6.4: (a) Median probability predictions by forecasters, the first-guess guidance, and the default linear decrease as a function of forecast time. (b) Median difference between the forecaster-issued predictions and the first-guess and linear decrease predictions as a function of forecast time. The shaded regions represent the 25th and 75th percentiles of the data distribution, calculated using 1000 bootstrapped samples.

the forecaster-issued and first-guess probability predictions of about 3% at all forecast times (Fig. 6.4b). Note that there were only 47 PHI warnings issued with a duration longer than 45 minutes.

This discrepancy between forecasters and the first-guess guidance at longer forecast times may be partially due to differences in what the probabilities being predicted actually represent. Specifically, the first-guess guidance was trained to predict the diagnostic ProbSevere probability of severe at a given forecast time. However, the section of the PHI prototype interface where forecasters modify these probabilities is labeled as "forecast confidence", or the forecaster-subjective probability that a storm will be severe at a given forecast time (Fig 2.2; Karstens et al. 2018; Kahneman and Tversky 1972; Brooks et al. 1992). While these two probability definitions can be interpreted in very different ways (e.g. high confidence of low probability, low confidence of high probability, etc.), anecdotal observations from the 2016 and 2017 PHI prototype experiments show that forecasters tend to blend these two probability schemes, such that the forecasters generally predicted probabilities consistent with the first-guess guidance at short to medium forecast times and probabilities consistent with subjective forecaster confidence at longer forecast times (decreasing with time). These observations are somewhat evident in the results shown in Fig. 6.4.

Both the forecaster-issued and first-guess probability predictions were found to be largely accurate when compared to the diagnostic ProbSevere probability corresponding to each forecast time of a given PHI warning plume. In general, forecasters tended to slightly over-predict severe probabilities at short to medium forecast times, and under-predict at medium to long forecast times. Similarly, the first-guess guidance typically predicted probabilities that were

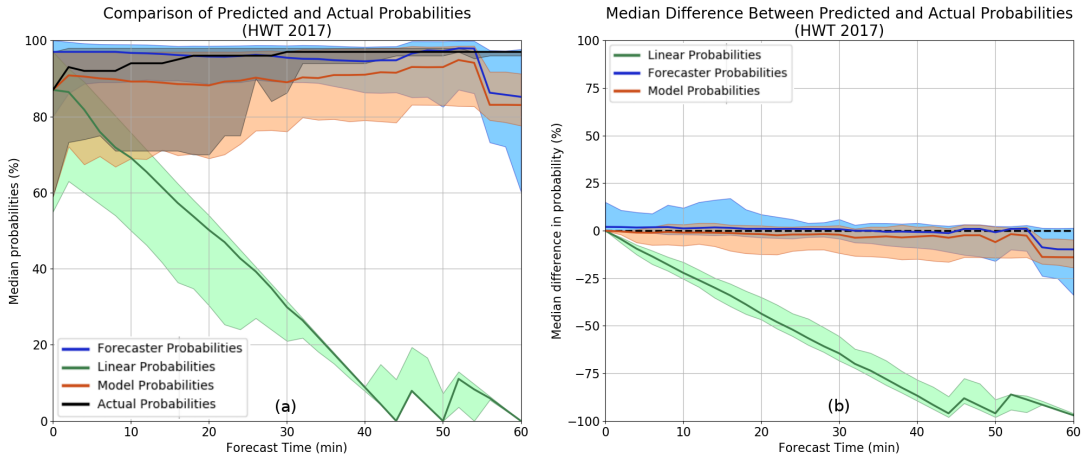

Figure 6.5: (a) As in Fig. 6.4 but with the true diagnostic ProbSevere probability. (b) As in Fig. 6.4, but for the difference between the actual ProbSevere probability and the forecaster and first-guess predictions.

lower than the true diagnostic ProbSevere probabilistic at all forecast times (Fig. 6.5a). Even so, these differences are shown to not be significant, as the median difference between the forecast and actual probabilities fell just along and slightly below the 0% difference line for forecast-issued and first-guess probability predictions respectively (Fig. 6.5b). In addition, the 0% difference line fell well within the 25th and 75th percentiles of each dataset, where each dataset contained 1000 bootstrapped samples as before. These results are quite interesting, as they show that the ensemble first-guess model performed well on most storms that forecasters issued warnings for. Additionally, forecasters were able to effectively utilize this new guidance as a first-guess and made adjustments that resulted in a net improvement in the predictions.

Interestingly, there was a notable discrepancy in the reliability of forecaster-issued and first-guess predictions, such that forecasters tended to significantly

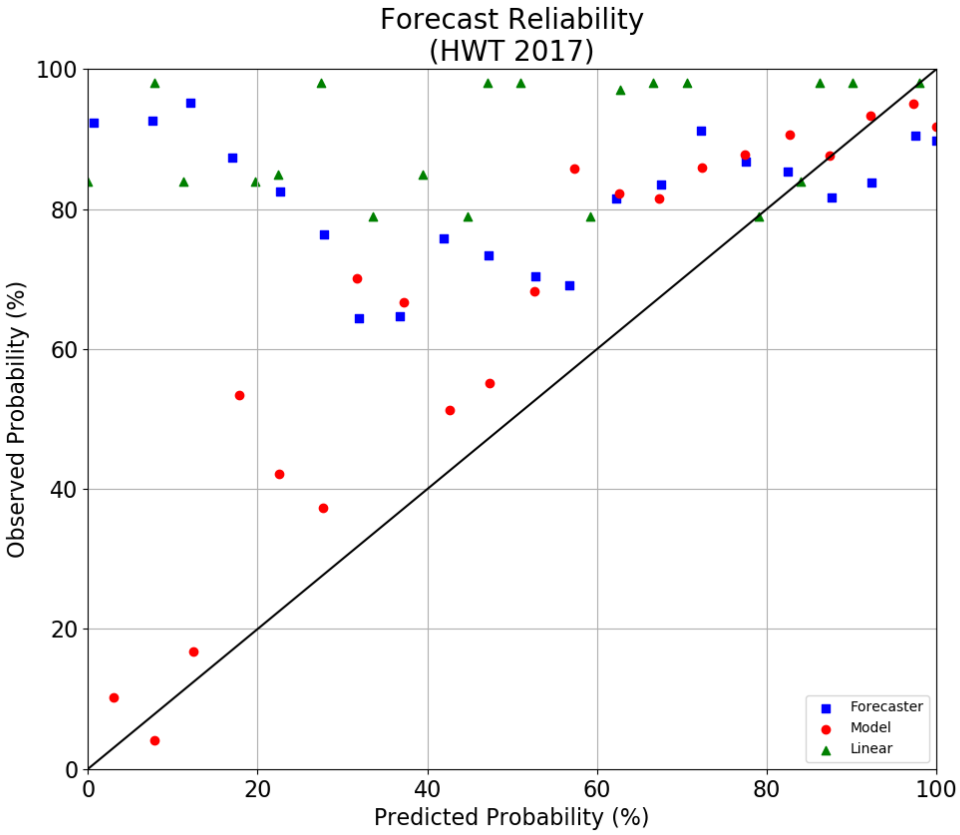

Figure 6.6: Reliability diagram for the forecaster-issued and first-guess probability predictions, as well as that of predictions from the default linear decrease.

under-forecast at lower predicted probabilities (Fig. 6.6). In contrast, probability predictions made by the first-guess ensemble generally followed the one-to-one line fairly closely, although it did tend to under-forecast somewhat in the medium probability spectrum. Again, this discrepancy can partially be explained by the differences in what the probabilities predicted by the forecasters and the model represent. As described previously, forecasters were noted to frequently decrease their probability predictions at longer forecast times, presumably corresponding to a decrease in their personal confidence that the storm would still be severe at those times. However, if the ProbSevere diagnostic probability did not decrease, then the forecasters lower confidence predictions would result in an under-forecast of the true severe probability. Note that there were very few warnings where either the forecasters or the first-guess guidance predicted a severe probability less than 30%. In fact, most predicted probabilities at any forecast time during the 2017 PHI prototype experiment were in the upper-probability spectrum, or from about 80% to 100% (Fig. 6.7). Forecasters generally issued more high-probability forecasts than the first-guess ensemble, whereas the ensemble predicted more mid-range probabilities. This is likely a reflection of the high-end events selected for the case studies used in the experiment, where storms often had impressive radar signatures that would lead to high confidence that severe weather was occurring. This is further reflected by the number of high diagnostic probabilities of severe predicted by the automated ProbSevere guidance, where nearly all storm objects with a PHI warning had a diagnostic probability $> 90\%$.

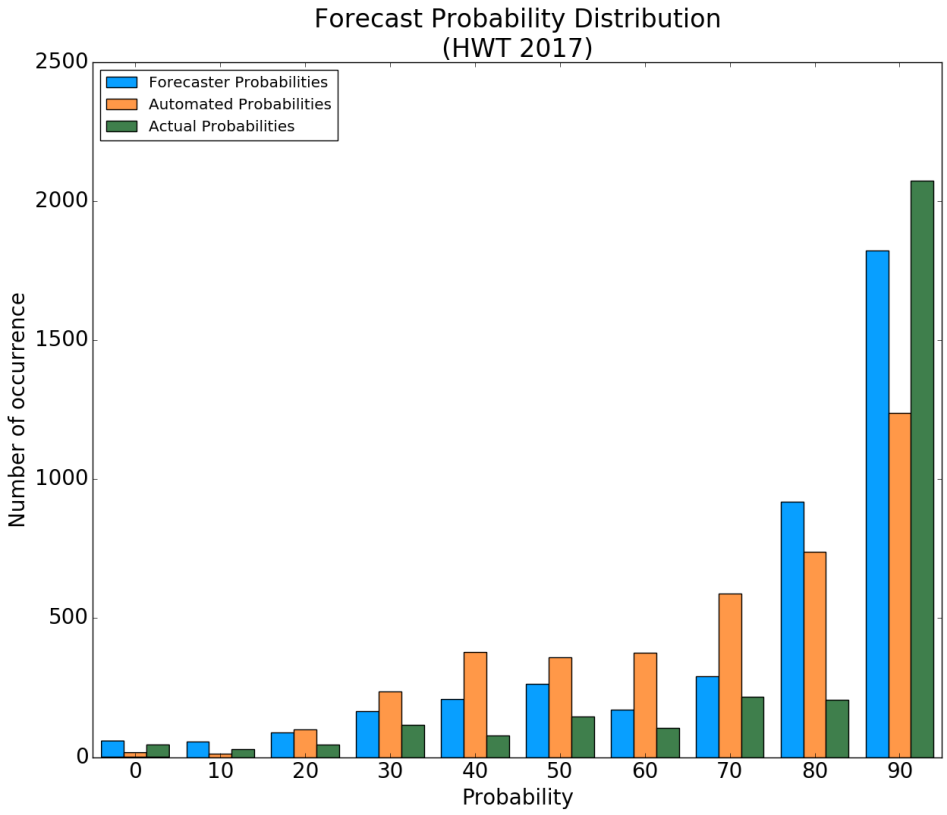

Figure 6.7: Distribution of actual and predicted probabilities at all forecast times during the 2017 PHI prototype experiment.

Chapter 7

Conclusions and Future Work

The first part of thesis examined the conditional verification statistics and predictability limitations of automated ProbSevere storm identification objects and their derived PHI warning plumes as compared to SBWs under the current warning paradigm. Less than 1% of all automated objects produced between April and June 2015 were associated with a storm that produced an LSR, but of those objects, about 80% of the derived warning plumes provided warning for the severe weather event at or near the time of occurrence. The automated guidance exhibited POD similar to NWS-issued SBWs during the analyzed time period, with 80% and 78% of all LSRs warned at the time of the event, respectively. Note that the FAA of the PHI warning plumes was not analyzed in this thesis. PHI plumes generally provided longer lead times for a severe weather event than SBWs, but this may be partially due to NWS guidance on the recommended maximum duration of SBWs. These results demonstrate the potential predictability limits of the automated PHI plumes, and suggest that the challenge for forecasters under the FACETs paradigm would be to correctly select which storms (ProbSevere objects) will produce an LSR at a rate that is comparable to SBWs to yield comparable verification metrics. In other words, forecasters can apply their current warning philosophy to PHI products and achieve similar or improved results.

Applying a spatial buffer to the ProbSevere objects and associated PHI plumes notably increased the fraction of plumes that verified at a given lead time, but only slightly improved the POD of the automated guidance at an LSR's time of occurrence. Instead, buffers were shown to potentially be more effective at improving lead time for a particular LSR by partially addressing model limitations when predicting the location of severe weather occurrence. In comparison, applying a spatial buffer to SBWs resulted in little change to POD or lead time. Furthermore, even unbuffered automated PHI plumes were shown to be spatially larger than current SBWs on average, and potentially less precise than the human-issued products. This presents a second challenge for forecasters to correctly select and buffer storms (objects) that will produce an LSR that falls just outside of the automated plume at the expense of anecdotally adding additional FAA to the implied warning polygon.

In addition, it was shown that the NWS forecaster participants in the 2017 PHI prototype experiment were able to achieve a POD of about 60% at an LSRs time of occurrence by utilizing and modifying the automated ProbSevere guidance. Furthermore, forecaster modifications resulted in improved performance of the warning plumes at longer lead times.

The second part of this thesis applied random forest, AdaBoost, gradient boosting, and elastic net machine learning techniques to produce an ensemble of models capable of accurately predicting how a storm's probability of producing severe weather will change with time. This ensemble was implemented in the 2017 PHI prototype experiment, and forecasters were asked to make probabilistic predictions using the new guidance. It was shown that forecasters used the initial ProbSevere probability guidance about 25% more frequently in the 2017 experiment compared to the one held in 2016, and forecasters only adjusted the

new probability trend guidance by a median of about 3% during the three week experiment. This suggests that forecasters generally trusted or agreed with the predictions being produced by the ensemble guidance, and indicates improved forecaster-guidance interdependence. Furthermore, both forecaster- and guidance-issued probability predictions were shown to be accurate compared to the diagnostic ProbSevere probabilities, although the ensemble predictions exhibited greater reliability than those issued by the forecasters.

Although the intent of this thesis is to leave forecasters and researchers with a better understanding of the automated PHI system, additional work is needed to further improve forecaster interdependence with the automated guidance. For example, future work may incorporate a larger sample size to better assess the verification of the ProbSevere model, while other research may apply grid-based methods, such as those described in Stumpf et al. (2015), to identify where improvements in FAA and plume probabilities are needed. Moreover, the ensemble probability trend predictions may benefit from additional training on a larger dataset and the inclusion of supplementary members trained using more complex learning techniques. Additional research (McGovern et al. 2018a; Jergensen et al. 2018) is currently underway to better identify and classify the convective characteristics of a storm (supercell, QLCS, etc.), which may lead to improved forecasts of a storms expected duration and the likelihood that it will produce severe weather during that period. These and other improvements to the current automated guidance are hypothesized to be necessary for a more effective establishment of trust between forecasters and automated guidance. Ultimately, the findings presented in this thesis demonstrate the necessity of human forecasters in the warning issuance process, and the benefits that

forecaster-guidance interdependence may provide to enhance the meteorological communitys ability to issue timely, life-saving information to the public.

Reference List

Barlow, R. E., 1972: Statistical inference under order restrictions; the theory and application of isotonic regression. Technical report.

Benjamin, S. G., D. Devenyi, T. Smirnova, S. S. Weygandt, J. M. Brown, S. Peckham, K. J. Brundage, T. L. Smith, G. A. Grell, and T. W. Schlatter, 2006: From the 13-km RUC to the Rapid Refresh. *12th Conf. on Aviation, Range, and Aerospace Meteorology*, Amer. Meteor. Soc., Atlanta, GA, 9.1.

Besson, L., 1904: Attempts at methodical forecasting of the weather. *Monthly Weather Review*, **32**, 311–313.

Breiman, L., 1984: *Classification and regression trees*. Routledge.

—, 1997: Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California at Berkeley.

—, 2001: Random forests. *Machine learning*, **45**, 5–32.

Brier, G. W., 1944: *Verification of a Forecaster's Confidence and the Use of Probability Statements in Weather Forecasting*. US Department of Commerce, Weather Bureau.

—, 1946: *A study of quantitative precipitation forecasting in the TVA basin*. Number 26, US Department of Commerce, Weather Bureau.

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

Brooks, H. E., C. A. D. III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Weather and Forecasting*, **7**, 120–132.

Casteel, M. A., 2016: Communicating increased risk: An empirical investigation of the national weather services impact-based warnings. *Weather, Climate, and Society*, **8**, 219–232.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Weather and Forecasting*, **29**, 639–653.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, D. T. Lindsey, L. Cronce, J. Gerth, B. Rodenkirch, J. Brunner, and C. Gravelle, 2018: The NOAA/CIMSS ProbSevere Model: Incorporation of total lightning and validation. *Weather and Forecasting*, **33**, 331–345.

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

Dalton, J., 1793: *Meteorological observations and essays*. Harrison & Crosfield.

DelSole, T., 2004: Predictability and information theory. Part I: Measures of predictability. *Journal of the Atmospheric Sciences*, **61**, 2425–2440.

Domingos, P. and M. Pazzani, 1997: On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*.

Drost, R., M. Casteel, J. Libarkin, S. Thomas, and M. Meister, 2016: Severe weather warning communication: Factors impacting audience attention and retention of information during tornado warnings. *Weather, Climate, and Society*, **8**, 361–372.

Drucker, H., 1997: Improving regressors using boosting techniques. *ICML*, volume 97, 107–115.

Dykstra, R. L. and T. Robertson, 1982: An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, 708–716.

Efron, B. and R. Tibshirani, 1986: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.

Elsner, J. B., L. E. Michaels, K. N. Scheitlin, and I. J. Elsner, 2013: The decreasing population bias in tornado reports across the central plains. *Weather, Climate, and Society*, **5**, 221–232.

Ferree, J., 2006: NOAA/National Weather Service's storm-based warnings. *23rd Conf. Severe Local Storms*, Amer. Meteor. Soc., St. Louis, MO, P11.6.

Freund, Y. and R. E. Schapire, 1997: A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, **55**, 119–139.

Freund, Y., R. E. Schapire, et al., 1996: Experiments with a new boosting algorithm. *Icml*, Bari, Italy, volume 96, 148–156.

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

—, 2002: Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, 367–378.

Gagne II, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *Journal of Atmospheric and Oceanic Technology*, **26**, 1341–1353.

Gagne II, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and Forecasting*, **32**, 1819–1840.

Géron, A., 2017: *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.

Hamill, T., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bulletin of the American Meteorological Society*, **81**, 2653–2664.

Hamill, T. M., C. Snyder, and J. S. Whitaker, 2003: Ensemble forecasts and the properties of flow-dependent analysis-error covariance singular vectors. *Monthly Weather Review*, **131**, 1741–1758.

Hansen, T. L., G. J. Stumpf, K. L. Manross, C. Golden, A. V. Bates, J. G. LaDue, C. Ling, D. M. Kingfield, T. Meyers, , and N. Hardin, 2017: FACETs - The 2016 Hazard Services Probabilistic Hazard Information (HS-PHI) experiment at the NOAA Hazardous Weather Testbed. *33rd Conf. on Evnironmental Information Processing Technologies*, Amer. Meteor. Soc., Seattle, WA, J9.2.

Harrison, D. R. and C. D. Karstens, 2017: A climatology of operational storm-based warnings: A geospatial analysis. *Weather and Forecasting*, **32**, 47–60.

Harrison, J. K., C. Ellis, C. McCoy, H. Sorensen, and K. Williams, 2014: Evaluation of the National Weather Service impact-based warning tool. *Ninth Symp. on Policy and Socio-Economic Research*, Amer. Meteor. Soc., Atlanta, GA, J5.3.

Herman, G. R., E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of storm prediction center convective outlooks. *Weather and Forecasting*, **33**, 161–184.

Hintz, J. L. and R. D. Nelson, 1998: Violin plots: A box plot-density trace synergism. *The American Statistician*, **52**, 181–184.

Hitchens, N. M. and H. E. Brooks, 2012: Evaluation of the storm prediction centers day 1 convective outlooks. *Weather and Forecasting*, **27**, 1580–1585.

Hoerl, A. E. and R. W. Kennard, 1970: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoffman, R. R., D. S. LaDue, J. G. Trafton, H. M. Mogil, and P. J. Roebber, 2017: *Minding the weather: How expert forecasters think*. MIT Press.

Hwang, Y., A. J. Clark, V. Lakshmanan, and S. E. Koch, 2015: Improved nowcasts by blending extrapolation and model forecasts. *Weather and Forecasting*, **30**, 1201–1217.

Jergensen, E., A. McGovern, C. Karstens, H. Obermeier, and T. Smith, 2018: Real-time and climatological storm classification using support vector machines. *17th Conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, Amer. Meteor. Soc., Austin, TX, 1.2.

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. W. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The storm cell identification and tracking algorithm: An enhanced wsr-88d algorithm. *Weather and Forecasting*, **13**, 263–276.

Kahneman, D. and A. Tversky, 1972: Subjective probability: A judgment of representativeness. *Cognitive psychology*, **3**, 430–454.

Karstens, C. D., J. James Correia, D. S. LaDue, J. Wolfe, T. C. Meyer, D. R. Harrison, J. L. Cintineo, K. M. Calhoun, T. M. Smith, A. E. Gerard, and L. P. Rothfusz, 2018: Development of a human-machine mix for forecasting severe convective events. *Weather and Forecasting*, In press.

Karstens, C. D., D. LaDue, J. C. Jr., K. M. Calhoun, T. Smith, C. Ling, T. C. Meyer, A. McGovern, R. A. Lagerquist, D. M. Kingfield, B. T. Smith, E. M. Leitman, J. L. Cintineo, J. P. Wolfe, A. Gerard, and L. P. Rothfusz, 2017: Prototyping a next-generation severe weather warning system for FACETs. *Seventh Conf. on Transition of Research to Operations*, Amer. Meteor. Soc., Seattle, WA, 8.1.

Karstens, C. D., G. Stumpf, C. Ling, L. Hua, D. Kingfield, T. M. Smith, J. C. Jr., K. Calhoun, K. Ortega, C. Melick, and L. P. Rothfusz, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 hazardous weather testbed. *Weather and Forecasting*, **30**, 1551–1570.

Kohavi, R. et al., 1995: A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, Montreal, Canada, volume 14, 1137–1145.

Kuncheva, L. I., 2006: On the optimality of naive bayes with dependent binary features. *Pattern Recognition Letters*, **27**, 830–837.

Lakshmanan, V., B. Herzog, and D. Kingfield, 2015: A method for extracting postevent storm tracks. *Journal of Applied Meteorology and Climatology*, **54**, 451–462.

Lakshmanan, V., R. Rabin, and V. DeBrunner, 2003: Multiscale storm identification and forecast. *Atmospheric research*, **67**, 367–380.

Lakshmanan, V., T. Smith, G. Stumpf, and K. Hondl, 2007: The warning decision support systemintegrated information. *Weather and Forecasting*, **22**, 596–612.

Lewis, J. M., 2005: Roots of ensemble forecasting. *Monthly Weather Review*, **133**, 1865–1885.

Lindell, M. K. and R. W. Perry, 2012: The protective action decision model: Theoretical modifications and additional evidence. *Risk Analysis*, **32**, 616–632.

Ling, C., J. J. James, S. M. Miran, G. J. Stumpf, T. L. Hansen, K. L. Manross, J. LaDue, A. V. Bates, C. Karstens, K. M. Calhoun, J. C. Jr., T. C. Meyer, A. Gerard, and L. Rothfusz, 2017: Forecasters' mental workload while issuing Probabilistic Hazard Information (PHI) during 2016 FACETs PHI Hazardous Weather Testbeds. *33rd Conf. on Evnironmental Information Processing Technologies*, Amer. Meteor. Soc., Seattle, WA, J9.3.

Losego, J., B. Montz, K. Galluppi, M. Hudson, P. Browning, K. Runk, and K. Harding, 2013: Evaluating the effectiveness of IBW. *Eight Symp. on Policy and Socio-Economical Research*, Amer. Meteor. Soc., Austin, TX, 226066.

McGovern, A., K. L. Elmore, D. J. G. II, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, **98**, 2073–2090.

McGovern, A., E. Jergensen, C. Karstens, H. Obermeier, and T. Smith, 2018a: Real-time and climatological storm classification using machine learning. *17th Conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, Amer. Meteor. Soc., Austin, TX, 1.1.

McGovern, A., C. Karstens, D. Harrison, and T. Smith, 2018b: Using machine learning to predict storm longevity in real time. *17th Conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, Amer. Meteor. Soc., Austin, TX, J44.1.

Menzel, W. P. and J. F. W. Purdom, 1994: Introducing goes-i: The first of a new generation of geostationary operational environmental satellites. *Bulletin of the American Meteorological Society*, **75**, 757–782.

Mileti, D. S. and J. H. Sorensen, 1990: Communication of emergency public warnings. *Landslides*, **1**, 52–70.

Murphy, A. H., 1998: The early history of probability forecasts: Some extensions and clarifications. *Weather and forecasting*, **13**, 5–15.

Nichols, W., 1890: The mathematical elements in the estimation of the signal service reports. *Amer. Meteor. J*, **6**, 386–392.

NOAA, 2007: Storm-based warnings team report. *NOAA Tech. Rep.*, 45 pp.
URL `http://www.nws.noaa.gov/sbwarnings/docs/Polygon_Report_Final.pdf`

—, 2011: The historic tornadoes of april 2011. *NWS Service Assessment*, 77 pp.
URL `https://verification.nws.noaa.gov/images/Sats2_uploadCollections/SignedReport/sa1018SignedReport.pdf`

NWS, 2009: Storm-based warnings fact sheet. *NOAA/NWS*, 1 pp.
URL `http://www.nws.noaa.gov/pa/files/storm%20based%20warning.pdf`

—, 2014: WFO severe weather products specification. *National Weather Service Instruction 10-511*, 35 pp.
URL `http://www.nws.noaa.gov/directives/sym/pd01005011curr.pdf`

—, 2016: National Weather Service performance management interactive product database (updated daily).
URL `https://verification/nws.noaa.gov/idb/`

Pavolonis, M. J., 2010a: Advances in extracting cloud composition information from spaceborne infrared radiances - A robust alternative to brightness temperatures. Part I: Theory. *Journal of Applied Meteorology and Climatology*, **49**, 1992–2012.

—, 2010b: GOES-R advanced baseline imager (ABI) algorithm theoretical basis document for cloud type and cloud phase. Version 2. *NOAA/NESDIS/Center for Satellite Applications and Research (STAR)*, 96 pp.
URL `http://www.goesr.gov/products/ATBDs/baseline/Cloud_CldType_v2.0_no_color.pdf`

Roebber, P. J., D. M. Schultz, and R. Romero, 2002: Synoptic regulation of the 3 may 1999 tornado outbreak. *Weather and Forecasting*, **17**, 399–429.

Rothfusz, L., C. Karstens, and D. Hilderbrand, 2014: Forecasting a continuum of environmental threats: Exploring next-generation forecasting of high impact weather. *Eos, Trans. Amer. Geophys. Union*, **95**, 325–326.

Saha, S. and H. M. van den Dool, 1988: A measure of the practical limit of predictability. *Monthly Weather Review*, **116**, 2522–2526.

Scott, R. H., 1869: On the work of the Meteorological Office, past and present. *The Royal Institution Library of Science, Earth Science*, **1**, 333–345.

Sen, P. K., 1968: Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association*, **63**, 1379–1389.

Sheynin, O. B., 1984: On the history of the statistical method in meteorology. *Archive for History of Exact Sciences*, **31**, 53–95.

Sieglaff, J. M., D. C. Hartung, W. F. Feltz, L. M. Cronce, and V. Lakshmanan, 2013: A satellite-based convective cloud object tracking and multipurpose data fusion tool with application to developing convection. *Journal of Atmospheric and Oceanic Technology*, **30**, 510–525.

Smith, T. M., V. Lakshmanan, G. J. Stumpf, K. L. Ortega, K. Hondl, K. Cooper, K. M. Calhoun, D. M. Kingfield, K. L. Manross, R. Toomey, and J. Brogden, 2016: Multi-radar multi-sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97**, 1617–1630.

SPC, 2015: Storm Prediction Center storm reports (updated daily). URL `http://spc.noaa.gov/climo/reports/`

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC bioinformatics*, **9**, 307.

Stumpf, G. J., D. Karstens, and L. P. Rothfusz, 2015: Probabilistic Hazard Information (PHI): Highlighting the benefits via new verification techniques for FACETs. *Third Conf. on Weather Warnings and Communications*, Amer. Meteor. Soc., Raleigh, NC, 5.7.

Theil, H., 1950: A rank-invariant method of linear and polynominal regression analysis (parts 1-3). *Ned. Akad. Wetensch. Proc. Ser. A*, volume 53, 1397–1412.

Thompson, R. L., C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Weather and Forecasting*, **22**, 102–115.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

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. *Weather and Forecasting*, **21**, 408–415.

Whitaker, J. S., X. Wei, and F. Vitart, 2006: Improving week-2 forecasts with multimodel reforecast ensembles. *Monthly Weather Review*, **134**, 2279–2284.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*, volume 100. Academic press.

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

Witt, A., M. D. Eilts, G. J. Stumpf, E. D. W. Mitchell, J. T. Johnson, and K. W. Thomas, 1998b: Evaluating the performance of wsr-88d severe storm detection algorithms. *Weather and Forecasting*, **13**, 513–518.

Zou, H. and T. Hastie, 2005: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.