ABSTRACT

FRANCISCO, DIANNA MARIE. Associating Severe Weather Hazards in High-Shear Low-CAPE Environments within the Southeastern United States via Statistical Methods. (Under the direction of Dr. Lian Xie).

One of the biggest challenges for NOAA National Weather Service (NWS) operational forecasters is minimizing the false alarm rate (FAR) of severe weather hazards. Current forecasting techniques have difficulty predicting severe events that occur in high-shear, low-CAPE (HSLC) convective environments due to many reasons, including shallower storm depth. HSLC convective environments are prevalent in the Southeastern United States (SEUS), especially during the cool season and overnight time periods. A large proportion of the HSLC tornadoes and severe wind speeds are reported within SEUS. Much of the HSLC severe weather research has focused on synoptic and mesoscale patterns using likely variables identified through physical explanations to predict the observed storm reports. This research found combinations of variables that are associated with severe weather hazards to uncover connections that may otherwise not be explored. To address current needs of our NWS collaborators, this research study focused on the predictability of tornadoes and/or severe wind speeds within HSLC environments by: 1) identifying statistically significant weighted predictors that are weakly correlated among themselves, 2) creating training datasets that minimize spatial and temporal correlation errors in probabilistic models, 3) producing interpretable probabilities of tornadoes and/or severe wind speeds, and 4) decreasing FAR of these severe weather hazards.

A severe weather statistical procedure (SWSP) was constructed using a combination of statistical techniques, including correlation between predictors, smallest deviance with Y, and logistic regression, to identify variables that can distinguish between event cases and null cases using a binary response variable (Y, where event = 1, null = 0). The SWSP was used on a dataset

that included NWS storm reports labeled as event cases and unverified county warning area (CWA) centroids labeled as null cases. The sample areas were 5x5 grid boxes, so the maximum and minimum values of each predictor in the dataset were evaluated. North American Regional Reanalysis (NARR) data were used to test the SWSP and train probabilistic models. The probabilistic models gave an interpretable probability of tornadoes and/or severe wind speeds in HSLC convective environments where a warning was issued. These results provided a proof of concept for the SWSP, which was also an quick analysis tool.

Results from a NARR probabilistic model predicted a 36.4% FAR of a severe weather hazard (at the optimal cutoff) in the dataset consisting of unverified NWS warnings as null cases. When the number of missed events decreases, the FAR increases, if evaluating the same datasets; the only way to improve both values simultaneously is to improve the training datasets (e.g., Chapter III) and/or statistical model development process (e.g., Chapter II). Statistical considerations and data collection steps were then taken to create more appropriate training datasets for the SWSP regarding the prediction of HSLC severe weather hazards. Higher resolution data from the Rapid Refresh (RAP) model was also used to produce probabilistic models, to test the application to operational forecasting. A RAP probabilistic model predicted a 28.4% FAR of a severe weather hazard in the dataset.

It should be noted, this research focused on reducing FAR because the dataset used in this study only included cases in which a severe thunderstorm or tornado warning was issued in an HSLC environment. The model developed in this study can be extended to predict un-warned tornadic cases in HSLC environment in the future when such cases with sufficiently large sample size become well-documented and available. Finally, it should also be noted that the cases

identified as FAR may also contain uncertainty as tornadic events could occasionally go undetected. Such uncertainty can be reduced by improving the case dataset.

Associating Severe Weather Hazards in High-Shear Low-CAPE Environments within the Southeastern United States via Statistical Methods

by
Dianna Marie Francisco

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Marine, Earth, and Atmospheric Sciences

Raleigh, North Carolina

2019

APPROVED BY:

_____	_____
Dr. Lian Xie Chair of Advisory Committee	Dr. Gary Lackmann
_____	_____
Dr. Emily Griffith	Dr. Joseph Guinness

DEDICATION

To Wash and Simon, my darling Maine Coon cats who kept me calm and entertained during the entirety of this dissertation work.

BIOGRAPHY

Dianna M. Francisco received her Bachelor of Science in Meteorology/Mathematics (double major) at the University of Miami, and then earned her Master of Science in Atmospheric Science at the University of Nevada-Reno while conducting research at the Desert Research Institute. Dianna then returned home to Raleigh to pursue a Ph.D. in Atmospheric Science at NC State University where she studied atmospheric environmental conditions that produce severe weather within the Southeastern United States. While at NC State University, Dianna took a couple of Statistics courses where she developed an idea for a new forecasting technique using clustering and logistic regression, and this idea later became the foundation of her dissertation research. She continued learning statistics on her own and with the help of a Statistics Ph.D. student (now Dr.) Marcela Alfaro-Cordoba and two professors in the Department of Statistics, Dr. Emily Griffith and Dr. Joe Guinness. These informative conversations helped shape the statistical procedure (SWSP) introduced in this dissertation. In addition to her recently developed affinity for Statistics, Dianna enjoys nature walks, mountain trips to view waterfalls, painting, yoga, and spending time with her family and rambunctious kitties.

ACKNOWLEDGMENTS

I would like to acknowledge the NOAA NWS CSTAR award NA14NWS4680013, which provided partial funding through research assistantships; the Department of Marine, Earth, and Atmospheric Sciences (MEAS) for funding through teaching assistantships; The Graduate School for funding through the Doctoral Dissertation Completion Grant program; the North Carolina State University (NCSU) Libraries for funding through the Peer Scholars Program; and the NCAR CISL for free data storage and run time on the Yellowstone and Cheyenne high-performance computers. I would also like to acknowledge Andy Dean and Keith Sherburn for providing the HSLC case list and NARR dataset (in Ch. II) used in this research.

I would like to show my appreciation for my advisor, Dr. Lian Xie, and committee members, Dr. Gary Lackmann, Dr. Emily Griffith, and Dr. Joe Guinness. Dr. Xie was patient and kind to me throughout my time in the Coastal Fluid Dynamics Lab (CFDL), and I appreciate his understanding and support when I needed to take time off for health reasons. Dr. Lackmann has helped me through tough situations, and I appreciate his support and attention to detail. Dr. Griffith gave me the hope and encouragement I needed to continue my research when I felt like giving up early on, and she helped me improve my Statistical ideas. Dr. Guinness stepped in after other committee members left NCSU and I appreciate his kindness, statistical support, and collaboration. Thank you to my former committee members Dr. Montse Fuentes (Dept. of Statistics) and Dr. Bin Liu for their insightful conversations before they left NCSU and for time spent on my preliminary written exams. Also, thank you to all my labmates throughout the years in the CFDL for their friendship and lovely gifts from China.

I appreciate the Collaborative Science, Technology, and Applied Research (CSTAR) grant collaborators, in the NOAA National Weather Service and at NCSU, for their comments

and feedback at the beginning stages of this research. This includes lead PI Dr. Matt Parker who gave feedback on this research from 2015-2018, and fellow graduate students on the CSTAR grant who were part of the CSTAR monthly conference calls, group meetings, and discussions. I also appreciate Dr. Marcela Alfaro-Cordoba for helping me learn the R language in the earlier stages of my research. Thank you to the MEAS staff (especially Laura Holland and Meredith Henry), MEAS faculty, my fellow MEAS graduate students, and the undergraduate students I taught during my tenure at NCSU for all the helpful conversations and encouragement along the way. I am fortunate to have had numerous professional development opportunities provided through The Graduate School, including the Teaching and Communication Certificate program with Dr. Vanessa Doriott Anderson and Colleen McKearney, and dissertation writing support with Dr. Mike Carter and fellow Ph.D. Candidates in the Doctoral Dissertation Completion Grant program. I am also grateful for the professors of my Statistics courses, Dr. Herle McGowan and Dr. Justin Post, for my education and sparked interest in Statistics.

I would also like to acknowledge my appreciation to the three counselors I saw over the years for anxiety as well as survivor group support at the Counseling Center, the Health Center doctors including Physical Therapy and Women's Health, and the Yoga and Zumba instructors at the campus gym for keeping me healthy. I will forever appreciate Dan Wiltsie for persistently supporting me every day and taking care of our kitties and me throughout my erratic work schedule. My family and friends outside of academia have also supported me throughout this program, including keeping me grounded and reminding me about what matters most.

I couldn't have completed this Ph.D. program without the support of all these wonderful people and programs mentioned above. Thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

One of the biggest challenges for NOAA National Weather Service (NWS) operational forecasters is minimizing false alarms and raising the probability of detection of severe weather hazards, including tornadoes and severe wind speeds (NOAA 2018). An atmospheric environment that produces a severe weather hazard (i.e., a severe storm environment) is considered an event case. A false alarm, or false positive (FP), is defined as an unverified severe event warning (e.g., tornado warning); in other words, a warning was issued for a severe event (e.g., tornado) and that severe event was not observed (Figure 1.1). The percentage of severe event warnings that were unverified is referred to as the false alarm rate (FAR).

$$FAR = \frac{FP}{\#\ of\ nulls}$$

A confirmed event, or true positive (TP), is defined as a productive severe event warning; in other words, a warning was issued for a severe event and that severe event occurred (Figure 1.1). The percentage of confirmed severe events that were issued a warning beforehand is referred to as the probability of detection (POD).

$$POD = \frac{TP}{\#\ of\ events}$$

The FAR and POD of a model can be visualized by looking at the receiver operating characteristic curve (ROC curve). The area under the ROC curve (AUROC) illustrates the skill (robustness) of model by plotting the true positive rate (POD) versus the false positive rate (FAR). When the AUROC is equal to 1, the model has a POD of 100% and a FAR of 0%, which

is a perfect model (Figure 1.2); for this reason, the AUROC is a common model evaluation tool for severe weather forecasting models.

Although tornado watches are partially based on the environmental conditions predicted in numerical weather prediction (NWP) models, current tornado warnings are issued primarily due to Doppler radar observations of the parent storm, which gave a national average tornado warning lead time of 13 minutes in a 2009 study (Stensrud et al. 2009). Because radars only detect the rotation of the parent storm (i.e., mesocyclone), and not the rotation of a forming tornado, this forecasting technique contributed to the national average false alarm rate of 75% (Markowski and Richardson 2009, Stensrud et al. 2009). Another large contribution to the FAR is due to tornadogenesis occurring in parent storms that are not supercellular, i.e., no mesocyclone is present. This is because supercellular severe weather hazards are more commonly studied as well as easier to predict; prediction of tornadogenesis in non-supercells has more uncertainty (Brotzge et al. 2013, Caruso and Davies 2005). Since the Stensrud et al. (2009) study, the false alarm rate has dropped by a few percentage points but concurrently with a decrease in average tornado warning lead time and decrease in probability of detection (NOAA 2018). One reason for the decrease in skill is due to changes in how tornado warnings are issued and verified by the NWS, which previously were based on county warning areas (CWAs). A new method using storm-based warnings, which are based solely on the areas impacted by the event, was implemented before the start of the 2008 fiscal year. This presented a greater challenge because the warned areas are smaller, making them more difficult for forecasters to "hit," according to NOAA (2018). It is possible that in CWAs with smaller and/or more concentrated populations, higher FAR values are influenced by underreporting of severe weather hazards rather than over-warning (Anderson-Frey et al. 2016).

Regarding tornadoes during the latest fiscal year (2017), the national averages reported were a POD of 58%, 72% FAR and a 9-minute lead time (NOAA 2018). As for severe thunderstorms (which include tornadic storms) during the latest fiscal year (2017), the national averages reported were a POD of 81%, 50% FAR and a 17.7-minute lead time (NOAA 2018). The NWS defines a severe thunderstorm as a storm that produces: at least one tornado, winds of at least 50 knots (58 mph), and/or hail at least 1" in diameter. As shown statistically, unless improved forecasting techniques are implemented into forecasting operations (i.e., models become more skillful), decreasing FAR will result in a decrease in POD (i.e., if probabilistic model skill does not change, it will have the same ROC curve, so follow down along the ROC curve in Figure 1.2). In general, increasing the POD of events while decreasing the FAR of tornado/severe thunderstorm warnings (i.e., increasing model skill) is likely to lower the average tornado/severe thunderstorm warning lead time as shown in previous NWS evaluations (NOAA 2018). Therefore, it is recommended that these three measures of predictive skill be considered simultaneously when developing new severe weather forecasting techniques.

Researchers still do not know all the variables and processes involved in tornadogenesis, especially in environments other than the vastly studied high-shear, high-CAPE (HSHC) environments that commonly occur in Tornado Alley, i.e., the United States Plains (Stensrud et al. 2009; Markowski et al. 1998; Markowski and Richardson 2009, 2014a). HSHC environments are associated with tornadic supercells (Thompson et al. 2003) as well as tornadic non-supercells (Brady and Szoke 1989, Wakimoto and Wilson 1989); however, high values of CAPE are not necessary for tornadogenesis (Wheatley and Trapp 2008). High-shear, low-CAPE (HSLC) environments produce a large proportion of the tornadoes and severe wind reports within the Southeastern United States (SEUS), especially during the cool season and overnight hours (Johns

et al. 1993, Guyer et al. 2006, Brotzge et al. 2011, Sherburn and Parker 2014). HSLC convective environments include multiple convective modes, for example, Quasi-Linear Convective Systems (QLCS) and supercells, which affect the predictability of the severe weather hazard (Smith et al. 2008, Smith et al. 2010, Smith et al. 2012, Thompson et al. 2012, Davis and Parker 2014). Current forecasting techniques have difficulty predicting severe events that occur in HSLC convective environments due to many reasons including shallower storm depth (Davis and Parker 2014) and the rapid destabilization of the environment within the three hours prior to severe convection (King et al. 2017). Therefore, the timing of severe events, the convective mode, and the relatively small width and depth of the HSLC severe storm cells (Davis and Parker 2014) add to the difficulty of forecasting the probability of severe weather hazards (storm reports). As an example of the magnitude of the HSLC forecasting challenge, Dean and Schneider (2012) showed that over a ten-year period (2003-2012) the NWS had substantially higher false alarm rates (i.e., unverified tornado warnings) and higher number of missed tornadoes within HSLC environments compared to HSHC environments.

This research study focuses on the forecasting needs of the NWS as part of a research-to-operations collaboration with the NWS Weather Forecasting Offices (WFOs) throughout the SEUS and elsewhere. The research goal is to generate and provide severe weather forecasting products that benefit the NWS forecasting operations regarding difficult forecasting situations, specifically for HSLC environments. Due to the current forecasting challenges identified by the NWS and severe weather researchers, this study focuses on HSLC environments, which turns our attention to variables and processes that are directly associated with CAPE or that compensate for the lack of instability (i.e., other lifting mechanisms). Because most severe events (i.e., tornadoes and severe wind speeds) generated in HSLC environments occur in the

SEUS region (Sherburn and Parker 2014) as well as numerous tornado related deaths (Dean and Schneider 2012), this dissertation research will focus on HSLC cases within the SEUS. Understanding how processes, including lifting mechanisms, produce a tornadic environment with low-to-moderate CAPE (i.e., HSLC) in the SEUS has been investigated more in depth during recent studies (e.g., Sherburn et al. 2016, King 2016, King et al. 2017, Blank 2017, Sherburn 2018).

Most of the Southeastern HSLC severe research has focused on synoptic and mesoscale patterns using likely variables that have been previously identified as important indicators of severe weather through physical explanations to predict the observed storm reports (e.g., Sherburn et al. 2016, King et al. 2017). This research uses an approach in a different order by first finding the combination of variables that are associated with NWS official storm reports. This somewhat backwards approach has been successfully used in multiple climate studies, which used climate indices as predictors (e.g., Lo et al. 2007, Chen et al. 2013, Alfaro-Cordoba 2017). It is important to note that most severe weather composite parameters contain highly correlated predictors and/or assign equal weights to each component in the composite parameter (e.g., Rasmussen and Blanchard 1998, Thompson et al. 2004, Sherburn and Parker 2014, Sherburn et al. 2016). Some severe weather composite parameters centered and scaled the variables (predictors) inside the composite parameter post hoc, which can decrease the reliability and overall performance due to altering the predictor weights and values subjectively. There are also severe weather composite parameters that altered (transformed) the comprised variables without physical reasoning (e.g., squaring the variable). Variable (predictor) transformations should only be used to make the predictor value (X) versus predicted outcome (e.g., Y in a linear model) relationship (scatter plot) more linear, and it should not be done to superficially increase

the significance of the predictor. It is possible that some of the currently used severe weather composite parameters show skill partly because they are highly correlated with a more relevant variable or physical explanation, and/or contain highly correlated variables in the calculation of the severe weather composite parameter. Highly correlated variables provide much of the same information in a predictive equation (i.e., little independent information). Since each additional variable (i.e., predictor) in the equation can introduce more error in the prediction, reducing the number of highly correlated variables could be beneficial.

In addition, current severe weather composite parameters provide a value (prediction) that is difficult to interpret. For example, the Modified SHERB (MOSH) and Modified SHERB – Effective (MOSHE), which may currently be the most skillful severe weather composite parameters for HSLC convective environments (Sherburn et al. 2016), indicate when a severe event is likely but do not provide a probability of the severe event occurring. The MOSH and MOSHE also do not have a clear threshold value that indicates when the parameter is predicting an event or a null case. Probabilities cannot be successfully interpreted from forecasting models that were not designed to produce probabilities of occurrence, and probabilistic equations that are designed for a more specific forecast situation (e.g., HSLC convective environment) will be more skillful than probabilistic models designed with a more general focus (e.g., any convective environment in any location). With these issues in mind, the severe weather statistical procedure (SWSP) uses advanced statistical techniques to provide probabilistic equations that specifically predict HSLC severe events (i.e., tornadoes and severe wind speeds) within the SEUS with assumptions (data limitations) in mind.

The case list used in this research was also used to develop the MOSH and MOSHE (Sherburn et al. 2016), and it is important to be aware of the limitations of this dataset (see

section 2.7 for further discussion). The definition of the null cases in this dataset are essentially based on the subjectivity of the forecaster issuing the severe warning and the lack of observed storm reports. The explicit definition of the null cases, which are defined in Sherburn and Parker (2014), are: "the initial latitude–longitude point [as determined by the NWS operational meteorologist] of a severe thunderstorm or tornado warning that was issued in an HSLC environment [defined in section 2.3] when there were no severe reports from the Storm Data archives in the corresponding CWA [NWS county warning area] throughout that convective day (1200–1200 UTC)." To clarify, this includes all severe weather hazards (i.e., wind speeds greater than 50 knots, EF0 and greater tornadoes, and hail) for the entire day. The CWA and adjacent CWAs were used in the null definition to limit the possibility that a storm report occurred within the sample area of the null cases (see section 2.4). Because a severe thunderstorm and/or tornado warning was issued for all null cases, it is safe to assume that all cases occurred in convective environments, which is why it is stated that the resulting probabilistic models discriminate between HSLC "convective environments." These specific event and null definitions allow us to focus on the bigger impact events (stronger severe wind speeds and tornadoes) in HSLC convective environments, which were difficult to forecast by experienced NWS operational forecasters. Therefore, the models that are produced using this case list discriminate between two environments which appeared to be severe convective environments but may have had slight differences that are difficult to forecast and/or unknown (statistically) significant differences.

The event cases in the dataset also have a specific definition: event cases are NWS official storm reports that were EF1 or greater tornado reports and severe wind speed reports (wind gust $\geq$ 65 knots). Since the case definitions used in this study depend on the storm reports existing in the NOAA Storm Data archive (NOAA NCEI 2019), it is important to know the

limitations of this data archive. This includes discrepancies in the reports due to poor characterization of the scope and magnitude of estimated severe wind speeds (e.g., Doswell et al. 2005, Trapp et al. 2006, Smith et al. 2013), complexities of the damage-to-wind speed relationship (e.g., Doswell et al. 2009), subjectivity in tornado report severity and count (e.g., Verbout et al. 2006), primary convective mode that produces severe wind speeds in region (e.g., Smith et al. 2013), population bias and time of day (e.g., King 1997, Brooks et al. 2003, Trapp et al. 2006), and spatial distance between observed reports and actual reports (e.g., Sobash et al. 2011). The incompleteness and uncertainties of the case list, due to using the Storm Data archive, limit what the models predict. With the case list used for the probabilistic models in this research, these models predict a probability of a severe weather hazard (wind gust $\geq$ 65 knots and/or EF1+ tornado) in HSLC convective environments where a severe warning was issued (which is subjective). This is based on a partial representation of severe storm reports that occurred in HSLC environments in the SEUS (i.e., the severe weather hazards that were reported), which affects the event and null case list. The spatial and temporal definitions depend on the training dataset; see section 4.4. For further discussion on the case list definitions and reasoning, see section 2.3.

Therefore, the MOSH, MOSHE, and probabilistic models created from this dataset (see sections 2.3 and 2.4) need to be used carefully with all data definitions and limitations in mind. This research focused on reducing FAR, because it did not address the increase in POD which needed a larger amount of missed cases in the dataset. To develop a severe weather probabilistic model that is not just based on NWS-warned cases requires developing an additional database that includes a substantial number of NWS missed events; that database would be able to address the POD forecasting issue. The specific interpretation of the probabilistic models is based on the

training dataset used to produce the model. This includes the case definitions (e.g., event and null definitions), gridded model used to choose the predictors and coefficients (e.g., RAP), predictor values (e.g., maximum value), sample size (e.g., sample radius), sample time (e.g., difference between observation and model time), and data type (e.g., analysis, forecast). These specifications will also determine the spatial and temporal specifications of the prediction. That is, the probability of the severe weather hazard is valid within # km and # hours; see section 4.4 for more details. Interpretation of the probabilistic models is discussed further throughout the dissertation, including sections 2.3, 2.7, 4.4, 5.3, 5.4, and 6.2.

It is known that the public, including emergency managers, would like to see interpretable probabilities that they can intuitively understand and communicate (as defined in Table 2.10). Atmospheric researchers, NWS forecasters, and social scientists have all been working together to address the concern of consistent and clear messaging of weather forecasts. This research is focused on contributing to that goal. Atmospheric researchers have been utilizing statistical algorithms for decades but the ability to apply complex mathematical calculations to big data is a new development within the last two decades. This includes utilizing statistical methods to train models to predict an outcome such as the probability of tornadoes. A good overview of predictor selection techniques for statistical modeling, regarding tornado prediction, was given by Marzban et al. (1999). This study also highlighted the difficulty in assigning predictive strengths (e.g., coefficients) to predictors within a statistical model, which is discussed further throughout Chapter II. Commonly used statistical methods in severe weather prediction studies are discussed further in section 2.2.

When testing statistical models, a common misunderstanding is that a significant p-value equates to skill, which is untrue. For example, a probabilistic model consisting of only

statistically significant predictors, that is predictors with significant p-values, may not be a skillful model (section 3.2). Multiple studies, such as Gigerenzer (2004) and Ziliak and McCloskey (2009), examined the use of statistical techniques throughout science and found that the majority of the surveyed scientists inaccurately reported conclusions due to their misunderstanding of the statistics. This issue, centered around the misunderstanding of statistical significance and p-value, can be seen in severe weather research publications as well, which is why alternative validation techniques (e.g., cross validation) are recommended (Neville 2001). The definition of a p-value is the probability of the observed data given that the null hypothesis H_0 is true, i.e., $p(D|H_0)$; a significance test does not provide a probability for a hypothesis, i.e., $p(H_0|D)$ (Gigerenzer 2004). For example, a p-value of 0.01 does not mean there is a 99% probability that the predictor significantly discriminates between the null and event convective environments.

Besides the issue of misinterpreting significance, it is also important to note the statistical considerations and mindful data collection steps that can be taken when creating training datasets for severe weather statistical models. The temporal and spatial spacings of the input data (training dataset) should be carefully considered when producing probabilistic models for operational forecasting. This can include use of higher resolution gridded (NWP) model data such as the 13 km Rapid Refresh (RAP) model, but more importantly, it involves collecting data at spatial and temporal spacings that minimize correlation errors in the probabilistic models while also successfully capturing the local storm environment (Chapter III). In return, this can improve the predictive skill of the probabilistic models that lead to higher POD and lower FAR values.

To address current needs of the NWS, this dissertation research focuses on improving the predictability of tornado and/or severe wind speeds within low-instability environments (HSLC) by working towards: 1) identifying statistically significant weighted predictors that are weakly correlated among themselves, 2) creating training datasets that minimize spatial and temporal correlation errors in probabilistic models, 3) producing interpretable probabilities of tornadoes and/or severe wind speeds (severe weather hazards), and 4) decreasing FAR of these severe weather hazards. As previously stated, this research focused on reducing FAR, because it did not address the increase in POD which needed a larger amount of missed cases in the dataset. To achieve these goals, a severe weather statistical procedure (SWSP) was created to analyze hundreds of meteorological variables and choose a small set of appropriately weighted predictors, with correlations less than 0.4 among the predictors, that significantly discriminate between severe and non-severe (unverified warnings) HSLC convective environments (see Chapter II). The SWSP was theoretically designed to produce probabilistic models that give interpretable probabilities of any severe weather hazard(s) within any type of convective environment and for any geographical region. The specificity of the probabilistic model generated by the SWSP depends on the training dataset used. For this research study, the training datasets will be specific to tornado and/or severe wind speed reports (severe weather hazards) within a HSLC environment (convective environment) throughout the SEUS (geographical region), so the SWSP has only been tested using these specifications. The resulting probabilistic models can be used alongside the current NWS forecasting products to help communicators express the severe weather hazard probability in a way that may be more understandable to the public compared to coverage probabilities.

This research focuses on high-shear, low-CAPE (HSLC) convective environments within the Southeastern United States (Chapter IV) due to the forecasting challenges mentioned previously and because of the research grant priorities; however, the methods discussed in Chapters II and III can be used to develop probabilistic models for other types of convective environments. The novelty of this dissertation research lies within: 1) the development of a unique statistical procedure (SWSP), which creates probabilistic models that give interpretable probabilities of severe weather hazards (Chapter II), 2) highlights of statistical considerations for probabilistic modeling of severe weather hazards (Chapter III), 3) data collection methods that create appropriate training datasets for the SWSP (Chapter III), 4) probabilistic models that discriminate between severe and non-severe (unverified warnings) HSLC convective environments in the near-term (Chapter IV), and 5) applications of these HSLC probabilistic models to current operational forecasting (Chapter V). Interpretations of the probabilistic models and future applications of this research are discussed in Chapter VI.

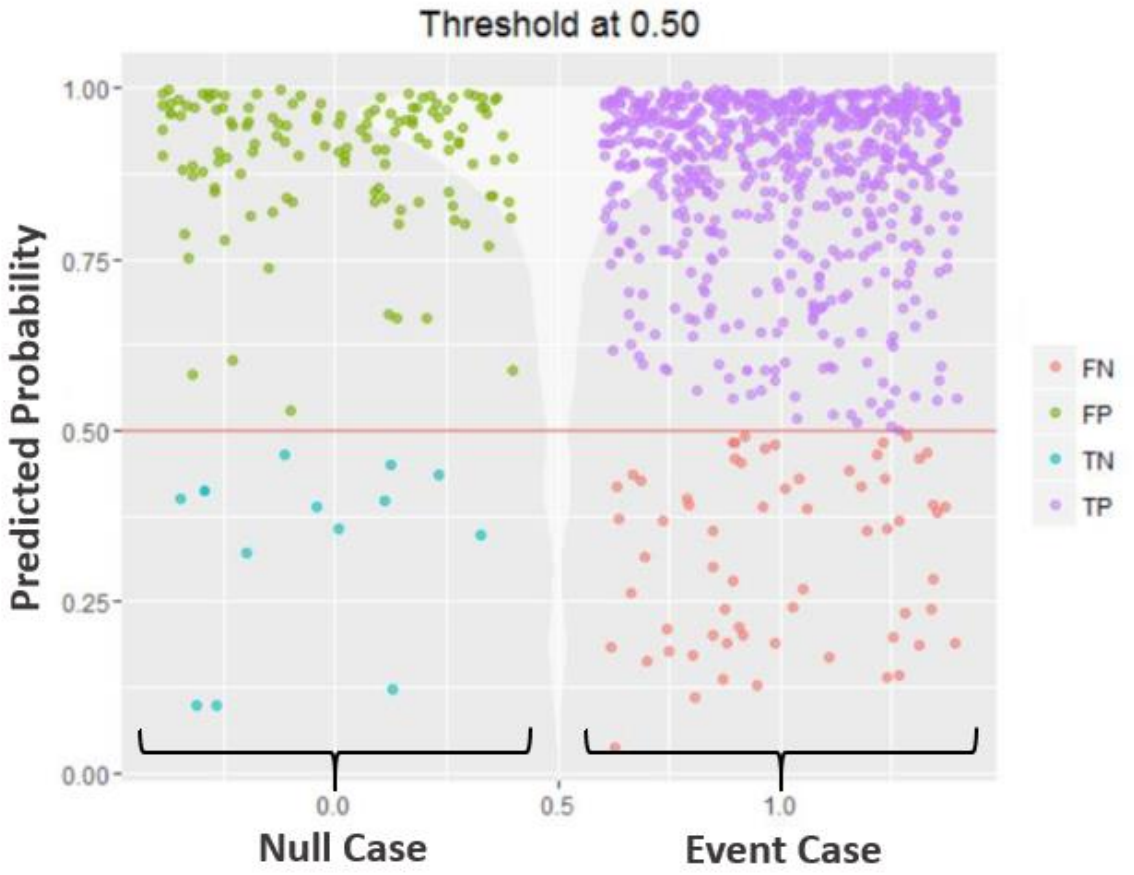


Figure 1.1. Visualization of a confusion matrix for demonstration only. The y-axis indicates the predicted probability (0-1) calculated by the model, and the x-axis shows the observed event case (1) or null case (0); dots are spread out to view more of the results, but they are technically on one vertical line either at 0 or at 1 on the x-axis. FN are false negative (i.e., missed event), FP are false positives (i.e., false alarm), TN are true negatives (i.e., correctly predicted null), and TP are true positives (i.e., correctly predicted event). The threshold (red line) is the optimal cutoff, which can be moved to determine the best ratio of false alarms (yellow-green dots) versus missed events (red dots). The optimal cutoff is chosen as the threshold with the lowest misclassification error. For example, if the threshold was raised to 0.625 in the graphic shown, missed events would increase by 40 while missed nulls (false alarms) would only decrease by 4, which increases the misclassification error by 36 cases (out of total number of cases). There can be models that only produce false alarms without any missed events; however, this may substantially increase the false alarm rate.

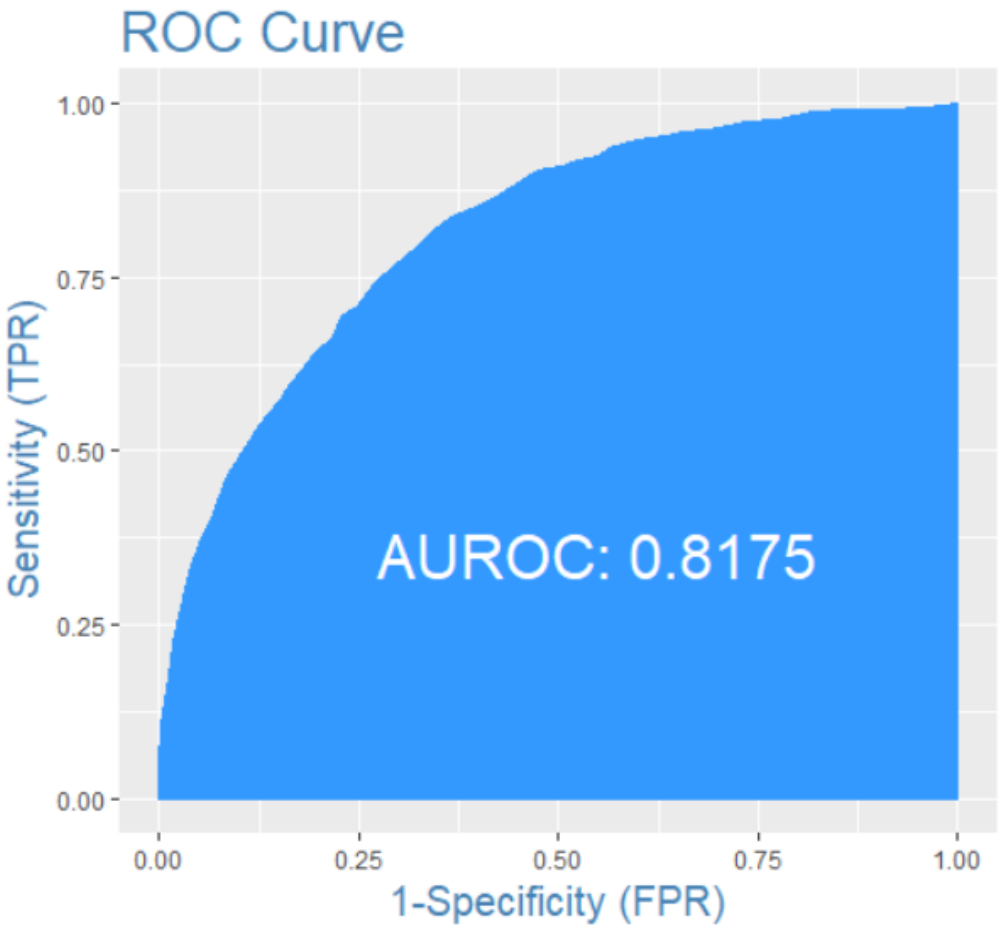


Figure 1.2. Visualization of the ROC curve and AUROC for a skillful probabilistic model. The receiver operating characteristic curve (ROC curve) illustrates the skill (robustness) of a logistic regression (binary) model by plotting the true positive rate (TPR) versus the false positive rate (FPR). The TPR is also referred to as the probability of detection (POD), and the FPR is also referred to as the false alarm rate (FAR); therefore, this graphic visualizes the POD and FAR of a model. Comparing with Figure 1.1, TPR equals TP divided by the number of events, and FPR equals FP divided by the number of nulls. The area under the ROC curve (AUROC) ranges from 0 to 1 and is used to determine skill in a statistical model with a value of 1 representing a perfectly skilled model (i.e., no error) and 0 representing a model with no skill. A model earning a value of 0.5 is considered to have skill equivalent to random chance; therefore, a model will be considered skillful if the AUROC is greater than 0.5. An AUROC is 1 when POD is 100% and FAR is 0%, which would fill in all 64 squares in the graphic shown. Note that if model skill does not change, it will have the same ROC curve; therefore, a decrease in FAR results in a decrease in POD.

CHAPTER II

Development of Statistical Procedure that Analyzes Highly Correlated Meteorological Data and Creates a Probabilistic Model

2.1. Introduction

The main goal of this research-to-operations dissertation study was to create probabilistic models that output interpretable probabilities of severe weather hazards on a regional-to-local scale for NWS operational forecasting. To attain this goal, a statistical procedure, referred to as the severe weather statistical procedure (SWSP), was developed to produce the probabilistic models. The SWSP was developed and finalized through a culmination of work conducted between 2014 and 2016. National Weather Service (NWS) needs and concerns were taken into consideration during the development of the SWSP. Throughout the development process, progress of the SWSP and preliminary results were shared with NWS collaborators in Weather Forecasting Offices throughout the Southeast and within the Storm Prediction Center (SPC).

The creation of the SWSP was motivated by the current challenges involved in developing equations that predict severe weather hazards (e.g., tornadoes and severe wind speeds) in the Southeastern United States (SEUS). The existing severe weather composite parameters and predictive equations typically only perform well in the region where the training data were focused (e.g., U.S. Plains). For example, the training dataset contained mostly severe weather hazard cases from the U.S. Plains region, and/or the training dataset was too general (i.e., not enough restrictive definitions) so the predictor values from the U.S. Plains cases dominated the outcome. Few severe weather forecasting equations have been developed with a focus on the SEUS region, especially for HSLC severe environmental conditions. There have

been studies that developed severe weather composite parameters or predictive equations that focused on severe weather development within SEUS, but these studies used larger spatial scale and temporal scale data and/or used techniques that do not take advantage of most of the predictive information within the training datasets. The SWSP was designed to produce probabilistic equations of severe weather hazards for any specified region and/or type of convective environment by inputting training datasets that are focused on those specifications. This creates more specific probabilistic equations, which can improve severe weather predictions. The SWSP code was written in a free open-source program (i.e., R), so the user can produce probabilistic models with any training datasets based on their specific needs. Minimizing the costs and computational resources necessary to run the SWSP and the subsequent probabilistic models was taken into consideration during the development of these techniques. NWS forecasters have limited time, money, and computational resources, so a procedure utilizing free software that quickly creates probabilistic equations with minimal user input was essential.

This chapter (II) focused on two research questions: 1) What is a new way to create a statistical procedure that can apply to various severe convective environments to produce an interpretable probability of a severe weather hazard?, and 2) can the implementation of statistically significant, weighted, and weakly intercorrelated predictors in probabilistic models provide a skillful interpretable probability of a severe weather hazard? As stated Chapter I, desired improvements included reducing the false alarm rate of severe weather hazards. Based on the current forecasting challenges (Chapter I), this research focused on producing interpretable probabilities of severe weather hazards within high-shear, low-CAPE (HSLC) convective environments. Specifically, this research focused on two severe weather hazards, tornado and

severe wind speed; hail and heavy rainfall were excluded because additional datasets would need to be produced to include additional predictors that represent microphysical and precipitation processes. Other reasons for the exclusion of hail are stated in Sherburn et al. (2016).

With the research goal and forecasting challenges in mind, the logistic regression method was chosen because it is the best method for producing interpretable probabilities of a severe weather hazard. Other statistical methods are useful in specific ways and may create probabilities (section 2.2); however, they do not produce interpretable probabilities as defined here: a probability that is directly associated with the odds that an event will occur. For example, a 30% probability corresponds to a 3 in 10 odds that an event will occur, which a logistic regression can directly state. This also means that when a "30% chance of tornado" is predicted, tornadoes will occur, on average, in 30% of the cases. For comparison, if the random forest method predicts a 30% probability, i.e., 30% of the forests (decision trees) predict an event will occur, it does not equate to a 30% probability the event will occur. Other statistical methods, including random forest, can be useful in weather prediction research, but the logistic regression method focuses on predicting events (e.g., severe weather hazards) accurately, which is the goal of this study. The main issues with the logistic regression, that the coefficients may change sign and/or become less interpretable or the model may be unstable due to large standard errors, occur when the independent variables (i.e., predictors) are highly correlated among themselves (i.e., collinear variables). However, this issue is mitigated by removing highly correlated predictors from the dataset before running the logistic regression, as discussed in section 2.5.

2.2. Background of Statistical Methods in Severe Weather Prediction

The decision of which statistical method to utilize was centered around the research goal of producing interpretable probabilities of severe weather hazards for NWS forecasting operations. This includes the constraints of choosing a statistical method that is quick and easy to calculate using a free statistical software (i.e., R). A method that works with minimal data storage is a bonus (e.g., creating one CSV file containing the necessary data to run the statistical method instead of holding onto the large gridded model output files). In previous research studies, numerous statistical methods were explored to examine convective environments and/or improve predictions of severe weather hazards. An overview of the common statistical methods used in weather prediction studies is given, including comments regarding the research goal (or problem). This overview is comprised of brief discussions on Principle Component Analysis (PCA), decision trees, random forest, linear regression, logistic regression, and the Bayesian approach.

The Principle Component Analysis (PCA) method finds uncorrelated derived variables that in succession maximize variance (Hotelling 1933, Jolliffe 2002). With a PCA, the variables in the dataset (i.e., original predictors) are transformed into principle components (PCs), which are uncorrelated and ordered so the first few PCs contain the most variation in all the variables in the dataset (Jolliffe 2002). Essentially, the PCA smooths the original predictor values by picking out the principle components (larger wavelengths) and leaving out the “noise” (smaller wavelengths) in the data. These PCs become the predictors for the statistical model. A model selection technique, for which there are many, then decides how many PCs to keep in the model; for the model to be skillful, it is important to avoid too many or too few PCs in the model.

Much development of the PCA and extended techniques, i.e., various techniques to summarize data that vary both spatially and temporally, have occurred within the field of atmospheric science (Jolliffe 2002). In atmospheric science, the eigenvectors that define the PCs are known as empirical orthogonal functions (EOF), and the values of the PCs are referred to as amplitude time series, coefficients, or EOF coefficients; however, using the term "coefficient" for a PC is technically a misnomer (Jolliffe 2002). PCA analysis and EOF analysis are sometimes treated as synonymous in atmospheric science, but they are distinguishable, and the term depends on how the eigenvectors are utilized. It is not recommended that the PCs be physically interpreted. A main reason for this disproval is because coefficients are estimated for the PCs, not for the original predictors (e.g., temperature). The PCA is dependent on the size and shape of the spatial domain over the observations, and it should only be used if the physically meaningful modes (of the atmosphere) are expected to maximize variance and/or be uncorrelated (Jolliffe 2002). Because the results of the PCA (or EOF) should not be physically interpreted, it was not a suitable method for this research problem.

Decision trees, a supervised machine learning method, are easy to interpret and useful for classification (e.g., automatically detecting the storm type based on radar data; Gagne et al. 2009). While decision tree results (e.g., storm type identification) can be useful for severe weather forecasting, the "decision" is categorical (e.g., yes or no) and does not provide a probability. However, a decision tree can show that a specific feature (i.e., predictor) in the model correctly predicts up to a (specified) percentage of the time, with the assumption that the learned decision tree predicts perfectly. This specified percentage is most useful as a comparison value, and this information can be used for variable selection by the user (not by the method) by choosing the predictors that are more important. To produce a probability using decision trees,

the random forest method is utilized. The random forest method is an ensemble consisting of hundreds of decision trees, which collectively produce a probability of an outcome (Ho 1995).

Random forest for classification (or regression) is a supervised machine learning method that has recently become popular in weather prediction studies (e.g., McGovern et al. 2011, Ahijevych et al. 2016, Herman and Schumacher 2018). Random forest with bootstrap aggregation is a popular method because it reduces overfitting the model compared to decision trees. The method is mainly used because it ranks the importance of the predictors, and then the user can select predictors that are most important for future use. It does this by searching for the best predictor among a random subset of predictors that are randomly selected with replacement to create each uniform subset, which later gives information on predictor importance. The relative importance of the predictor is based on their location in the decision tree; the most important predictors are at the top of the tree. The random forest method is also popular because it does not require predictors to have a linear (or monotonic) relationship with the response variable (Y).

However, random forest, like other machine learning methods, does not calculate coefficients. The random forest empirically estimates the prediction of occurrence; therefore, there is not an interpretable ability of inference on the predictors. Since the random forest for classification cannot produce standard errors for either the prediction or coefficients, there are no confidence intervals for either the response variable or the predictors, which makes inference difficult. Other versions of the random forest method can produce an empirical confidence interval for the prediction, but this is not methodologically valid. Subsequently, the percentage given at the end of a random forest computation is an empirical count of the decision trees themselves, so the resulting probability is not interpretable, and there is not a logical way to

assess variability. There is also no model selection with this method and it does not give the significance of predictors in the model; therefore, this method includes predictors in the model regardless of the significance (personal communications with Amanda Muyskens, Ph.D. in Statistics).

Regarding operational use, the random forest must be recalculated each time a prediction needs to be made because a shareable prediction equation is not produced. Once a random forest is built, the predictor values are then plugged into the random forest to produce a prediction. Therefore, the entire random forest code needs to be shared with forecasters to receive the same results. Because the random forest method calculates hundreds of decision trees to produce a skillful prediction, it currently takes days and/or substantial computational resources to produce a result (personal communications with Amanda Muyskens, Ph.D. in Statistics). This may or may not be a disadvantage depending on the desired outcome. One study found that the random forest can run in a shorter time compared to higher resolution NWP models (Ahijevych et al. 2016). In all, the random forest method is currently ineffective for real-time severe weather predictions, i.e., near-term operational forecasting.

Methods that are currently effective for real-time predictions include linear regression methods. Linear regression methods may be the oldest statistical methods (in this overview list) used in weather prediction, going back to the creation of the model output statistics (MOS) in the early 1970s, which are based on multiple linear regression (Glahn and Lowry 1972). This method is mainly used to model a continuous response variable (Y) to correct systematic NWP model bias (i.e., mean error). This is done by calculating the average of the differences between the forecast value and the observation value (i.e., bias) for a variable (e.g., temperature), and then using that bias to produce a prediction of the variable value by “nudging” the NWP predicted

value (i.e., forecast value) closer to the observation value. MOS is also used to produce probabilistic forecasts based on the deterministic (NWP) model output. In general, linear regression methods are chosen because of their ability to interpret confidence intervals, which allows these methods to make inference. In other words, these methods take a result (conclusion) from the subset of the population (i.e., training dataset) and infer that result to the entire population (i.e., all cases including those not in the dataset). Therefore, linear regression allows interpretability of the prediction; this does not mean an interpretable probability as defined in Table 2.10.

Like linear regression methods, the logistic regression method is quick and easy to calculate. Also similar to linear regression, the logistic regression is able to make inference; however, it is used to model a binary response variable (Y) and thus the outcome is an interpretable probability (Cox 1958). In weather prediction, logistic regression is specifically used for developing probabilistic guidance and forecasts of weather phenomenon (e.g., Coniglio et al. 2007, Mecikalski et al. 2015). A logistic regression (i.e., logit regression) uses a logistic function to model a binary response variable (i.e., Y where Y = 0 or 1), which gives a probability varying from 0 to 1 (i.e., 0% to 100%) as the result (Cox 1958). The logistic function converts the log-odds to probability, and the unit of measurement for the log-odds scale is called a logit (i.e., logistic unit). The logistic function has also been used with other statistical methods in severe weather prediction studies, such as the activation function for layers in a neural network (e.g., Marzban and Stumpf 1996). To use the logistic regression, there is an assumption of a linear relationship between the logit of the outcome (logit(Y)) and each predictor (X); in other words, logit(p) = log(p/(1-p)) where p is the probability of an event occurring. There is no assumption of a direct linear relationship between the response variable (Y) and each predictor

(X), which allows for predictors with non-linear relationships with Y in the training dataset. Specifically, there is a sigmoidal (s-shaped) relationship between the probability (p) and predictor (X). Therefore, the logit of the probability, and not the probability itself, follows a linear model. For clarity, throughout this dissertation the independent variables are referred to as predictors and the dependent variable is called the response variable.

The general benefits of the logistic regression method are: 1) it will fit a linear regression to the dataset to produce a probabilistic equation within one minute of computational time; 2) the probabilistic equation can be used to calculate future predictions without rerunning the statistical methods; 3) it prints coefficients and p-values (significance) for each predictor; 4) it is less prone to overfitting, which means the prediction performance on a new dataset is similar to the performance on the training dataset; and 5) the prediction is interpretable, such that relationships between the predictors (X_n) and response variable (Y) can be easily diagnosed. The main disadvantages of the logistic regression method are: 1) it may not predict complex non-linear data as well as other models; and 2) when the independent variables (i.e., predictors) are highly correlated (i.e., collinear variables), the coefficients may change sign and/or become less interpretable and the model may be unstable due to large standard errors. However, these issues can be mitigated through careful data collection and data processing steps when creating the training dataset for the logistic regression input.

Statistical methods discussed above could also be used in a Bayesian framework, but it is more complicated and computationally expensive (e.g., Röpnack 2013) than the "classic" frequentist model approach. The Bayesian model approach is computational expensive due to estimating a distribution for every variable (predictor) in the dataset (Berger 1985). To put it simply, the Bayesian approach uses a distribution (range) of values for each predictor (e.g.,

temperature from 2 to 3 degrees) instead of a single value (e.g., temperature of 2.5 degrees). Therefore, the outcome from the Bayesian approach is a distribution of values (e.g., probabilities from 0.2 to 0.4) instead of a single value (e.g., probability of 0.3). Prior information is needed for this approach to work properly, otherwise the results are not much different from the frequentist approach (i.e., using a single value). The Bayesian approach is a more complicated approach, but it is better at quantifying uncertainties and assessing ensemble forecasts, which is helpful in severe weather prediction. With a Bayesian statistical method, at least 10,000 iterations of sampling from the prior information is conducted, i.e., sampling over 10,000 times from one distribution of values per predictor (personal communications with Amanda Muyskens, Ph.D. in Statistics). The Bayesian approach essentially combines the prior guesses and dataset together, so the solution is somewhere between them. Since strong prior information is not currently available regarding the research problem, uninformative priors would have to be utilized, which would not help the analysis (i.e., solution would be close to the frequentist approach). For this reason, as well as the computational intensity, utilizing a Bayesian approach is outside the scope of this research.

When reviewing previous statistical methods that produce predictive equations of severe weather environments/hazards, the following observations were noted, with this dissertation's research goal in mind:

- Most methods do not produce an interpretable probability. Instead these predictive equations have "severe thresholds" (e.g., a composite parameter value greater than 1.0 is a "high probability" of an event occurring), or give a coverage probability.
- Most methods are not developed for lower instability environments (i.e., HSLC), and may not work well in this type of severe environment.

- Some methods started with a short list (< 12) of commonly used pre-selected predictors to produce the predictive equation, therefore new predictors were not sought out or tested.

- Some methods started with a large list of predictors but tested each predictor separately, and not in combinations, before choosing predictors for the equation.

- Some methods started with a large list of predictors and compared all possibilities at the same time but did not test/rank the predictors by statistical significance.

- Some methods started with a large list of predictors but pre-processed the predictor values with PCA before choosing predictors, which is not recommended for multiple reasons including "losing" predictive information.

- Some methods do not consider the multicollinearity of the predictors or possible spatial and/or temporal correlation errors.

- Some methods, including current severe weather composite parameters, have at least one predictor within the equation that highly correlates with another predictor in the equation.

- Some methods do not assign coefficients to weigh each predictor based on predictor importance, but instead all predictors have the same weight and are usually normalized.

- Some methods take days to produce a prediction, which is not useful for real-time operational forecasting.

2.3. HSLC Definitions & Case List

This research study was part of the North Carolina State University's 2014-2018 NWS CSTAR grant that focused on improving the predictability of tornadoes and severe wind speeds within HSLC environments. Another research study within the grant (Sherburn et al. 2016)

utilized a case list of tornado and severe wind reports that occurred in HSLC environments within the Southeastern United States (SEUS). Because the case list was specifically developed to study HSLC environments in the Southeastern United States, it was beneficial to use this case list for this study. A HSLC environment is defined when the atmospheric environment contains 0-6 km bulk shear vector magnitude ≥ 18 m/s, surface-based CAPE (SBCAPE) ≤ 500 J/kg, and most-unstable CAPE (MUCAPE) ≤ 1000 J/kg (Sherburn and Parker 2014, Sherburn et al. 2016). For this research, the SEUS region is defined as approximately 25° to 40° North (latitude), and 95° to 75° West (longitude) (Sherburn and Parker 2014).

The case list utilized by Sherburn et al. (2016), an expanded case list from Sherburn and Parker (2014), was used in this dissertation research as the "master case list" from which relevant Southeastern HSLC cases from 2006 to 2014 were chosen for the statistical model training datasets. This case list was provided to Sherburn et al. (2016) by Andy Dean at the NWS SPC. The master case list consisted of 2063 total cases, which were divided into 1326 events cases and 737 null cases (Figure 2.3). The event cases comprised of 605 severe wind speed reports and 720 tornado reports, and the null cases comprised of 593 severe thunderstorm warnings and 145 tornado warnings. The cases are discussed further below and in Figures 2.4 and 2.5. All cases were crosschecked with the NWS National Hurricane Center archives for the 2006 to 2014 hurricane seasons. This crosscheck ensured that there were no cases in the master case list associated with landfalling tropical cyclones.

This master case list included official storm reports from the NWS for all Southeastern HSLC cases from 2006 to 2014, which were labeled as "event cases." Specifically, these NWS official storm reports were EF1 or greater tornado reports and severe wind speed reports (wind gust ≥ 65 knots). EF0 tornado reports and severe wind speed reports less than 65 knots were

excluded from the case list because these reports have greater uncertainties on its classification (i.e., is it a tornado report or severe wind speed report) due to numerous reasons, including poorly estimated wind speeds within this range (e.g., Doswell et al. 2005), as discussed further in Sherburn (2018). Although this excluded a portion of the storm reports, it helped distinguish between non-tornadic "severe wind speed" reports from "tornado" wind reports, which allowed us to develop statistical models for a specific severe weather hazard (e.g., tornado). It is recommended to include the EF0 tornado reports in the case list when developing a general severe weather statistical model (i.e., tornadoes and severe wind speeds), but the EF0 cases were not available in the master case list.

The master case list also included "null cases," which are Southeastern HSLC cases with no verified severe storm reports within the CWA (i.e., NWS defined county warning area) or adjacent CWA throughout the convective day (i.e., 12Z to 12Z), and a severe thunderstorm warning or tornado warning was issued within the CWA (of null location) by the NWS. To clarify, this includes all severe weather hazards (i.e., wind speeds greater than 50 knots, EF0 and greater tornadoes, and hail) for the entire day. The CWA and adjacent CWAs were used in the null definition to limit the possibility that a storm report occurred within the sample area of the null cases (see section 2.4). The location of the null was defined as the initial latitude-longitude point of the warning issued by the NWS forecaster, which was essentially the centroid of the warning area. After the 2007 fiscal year, the NWS has used storm-based warnings that are smaller polygons based solely on the areas impacted by the warning and event (NOAA 2018).

The null cases are assumed to be true null cases, which assume that no severe weather hazard occurred within the sample area. This is also how null cases were classified in previous studies (e.g., Sherburn and Parker 2014, Sherburn et al. 2016). All conclusions are based on this

assumption. Since a severe thunderstorm and/or tornado warning was issued for all null cases, it is safe to assume that all cases occurred in convective environments, which is why it is stated that the resulting probabilistic models discriminate between HSLC "convective environments." For the needs of this research study, severe thunderstorm warnings were labeled as a "severe wind speed" null, and tornado warnings were labeled as a "tornado" null, when the datasets were separated by severe weather hazard type. Because of the specific event and null definitions, this allows us to focus on the bigger impact events (stronger severe wind speeds and tornadoes) in HSLC convective environments which were difficult to forecast by experienced NWS operational forecasters. Therefore, the models that are produced using this case list discriminate between two HSLC convective environments which appeared to be severe convective environments but may have had slight differences that are difficult to forecast and/or unknown (statistically) significant differences. Previous studies have developed ways to address the observation bias of storm reports; for example, Widen et al. (2013) used the average tornado report density as a function of distance from city/town to correct statewide tornado probabilities. Another study, Elsner et al. (2013), used a statistical model for report density as a function of distance from the nearest city center to evaluated population bias on tornado reports; this study can help researchers develop more realistic spatial tornado climatologies. These studies can help correct the tornado counts, but cannot estimate the location and time of a tornado that was not observed; therefore, correction methods like these above were not utilized in this dissertation research which utilized observed cases.

Since the case definitions used in this study depend on the storm reports existing in the NOAA Storm Data archive (NOAA NCEI 2019), it is important to know the limitations of this data archive. This includes discrepancies in the reports due to poor characterization of the scope

and magnitude of estimated severe wind speeds (e.g., Doswell et al. 2005, Trapp et al. 2006, Smith et al. 2013), complexities of the damage-to-wind speed relationship (e.g., Doswell et al. 2009), subjectivity in tornado report severity and count (e.g., Verbout et al. 2006), primary convective mode that produces severe wind speeds in region (e.g., Smith et al. 2013), population bias and time of day (e.g., King 1997, Brooks et al. 2003, Trapp et al. 2006), and spatial distance between observed reports and actual reports (e.g., Sobash et al. 2011). The incompleteness and uncertainties of the case list, due to using the Storm Data archive, limit what the models predict. With the case list used for the probabilistic models in this research, these models predict a probability of a severe weather hazard (wind gust ≥ 65 knots and/or EF1+ tornado) in HSLC convective environments where a severe warning was issued (which is subjective). This is based on a partial representation of severe storm reports that occurred in HSLC environments in the SEUS (i.e., the severe weather hazards that were reported), which affects the event and null case list. The spatial and temporal definitions depend on the training dataset; see section 4.4.

The probabilistic models created from this dataset need to be used carefully with these definitions in mind; see section 2.7 for information on data limitations. Also, this research focused on reducing FAR, because it did not address the increase in POD which needed a larger amount of missed cases in the dataset. To develop a severe weather probabilistic model that is not just based on NWS-warned cases requires developing an additional database that includes a substantial number of NWS missed events; that database would be able to address the POD forecasting issue.

2.4. HSLC Numerical Dataset

The Sherburn et al. (2016) dataset was used as the example dataset for the creation of the SWSP. This dataset was chosen as the best estimate of the regional environment (i.e., closest to observations), which is available over a time period that provides enough cases to build a robust statistical model. A benefit to using this already developed dataset was the ability to compare the results of the procedure with the Sherburn et al. (2016) results, which helped evaluate the performance of the SWSP. The values in the Sherburn et al. (2016) dataset were taken from the North American Regional Reanalysis (NARR), using a sample area of exactly 5x5 grid boxes (i.e., 162x162 km^2). The 5x5 grid boxes sample area included the grid box encompassing the case location (the center grid box) and all grid boxes surrounding the center grid box. The NARR uses the NCEP Eta Model, with 32 km horizontal grid spacing and 45 vertical layers, together with the Regional Data Assimilation System (RDAS) that assimilates the model variables (Mesinger et al. 2006, NARR 2016). NARR data is output at 29 vertical pressure levels, and it has a larger vertical spacing (i.e., 50 hPa instead of 25 hPa) in the output between 700 hPa and 300 hPa. The NARR was the highest resolution reanalysis dataset available with enough cases to build a robust statistical model (at the start of this research). The provided NARR dataset also included calculated variables (predictors) using the NARR output, such as severe weather composite parameters (e.g., MOSHE).

Some of the severe weather composite parameters that are currently used in severe weather forecasting are the Supercell Composite Parameter (SCP), Significant Tornado Parameter (STP), Vorticity Generation Parameter (VGP), and the Energy-Helicity Index (EHI), as well as the Severe Hazards in Environments with Reduced Buoyancy (SHERB), Effective SHERB (SHERBE), Modified SHERB (MOSH), and Effective MOSH (MOSHE), which are

specific to HSLC convective environments. (Note that these are all linear models.) The SCP discriminates between supercells and non-supercells (i.e., identifies supercell potential); more specifically, SCP significantly discriminates between surface-based supercells and surface-based, discrete non-supercells. The STP discriminates between tornadic and non-tornadic supercells. The mathematical definitions of SCP and STP are shown in Thompson et al. (2003) with modifications shown in Thompson et al. (2004). The modifications of SCP include an effective bulk shear term, and the modifications to STP include effective storm-relative helicity (SRH) and convective inhibition (CIN) terms. Every term in both the STP and SCP (i.e., every individual predictor within the composite predictor) is normalized and some terms are also based on parcel CAPE and CIN constraints. The VGP estimates the tilting and stretching of the horizontal vorticity by a thunderstorm updraft. The original VGP equation is a partial derivative that essentially equals the mean shear (i.e., hodograph length divided by depth) multiplied by the square root of CAPE (Rasmussen and Blanchard 1998). Above a value of 0.2, an increasing VGP suggests an increasing potential for tornadic storms. The EHI combines CAPE and shear (via storm-relative helicity) predictors to predict the potential for supercell storms, with values larger than 1.0 indicating potential for supercells and values larger than 2.0 indicating a “large probability of supercells” (Rasmussen and Blanchard 1998). The mathematical definitions of VGP and EHI are shown in Rasmussen and Blanchard (1998) with updates shown in Rasmussen (2003). Specifically, the 0-3 km VGP and 0-3 km EHI were included in the HSLC dataset. The SHERB, SHERBE, MOSH, and MOSHE are severe weather composite parameters that were created using Heidke Skill Scores (i.e., HSS; NOAA NWS 2018) to forecast HSLC severe environments (i.e., using the “all” dataset that included all cases). These composite parameters are comprised of terms (predictors) that earned high HSS (e.g., 0-1.5 km bulk shear vector

magnitude), and all predictors are normalized. The MOSH is a modified version of the SHERB, and the MOSHE is a modified version of the SHERBE. The mathematical definitions of SHERB(E) and MOSH(E) are shown in Sherburn and Parker (2014) and Sherburn et al. (2016). For more information on all the calculated variables (predictors) within the HSLC dataset, see Sherburn (2018).

This NARR dataset from Sherburn et al. (2016) was received as multiple datasets, including: maximum value (within a 162x162 km^2 area centered over the case location), minimum value (within a 162x162 km^2 area centered over the case location), 3-hour difference of maximum (i.e., value at report time minus value at 3hr prior), and 3-hour difference of minimum (i.e., value at report time minus value at 3hr prior). The last two datasets were specifically requested and not included in the Sherburn et al. (2016) paper. These last two datasets were requested based on a result in King et al. (2017) – the rapid destabilization of the environment within the three hours prior to severe convection – to see if environmental changes after convective initiation could be significantly related to the occurrence of a tornado and/or severe wind speed. The list of variables (predictors) within the datasets are shown in Table 2.1; there are 262 variables in the list. For additional study, all datasets were then separated into: EF1+ tornado events and all tornado nulls ("tornado" dataset), EF2+ tornado events and all tornado nulls, and severe wind speeds and severe thunderstorm nulls ("severe wind speed" dataset); there was also an "all" dataset that included all cases and nulls. Each dataset included both event cases (labeled as Y=1) and null cases (labeled as Y=0) to allow for significant difference testing and construction of probabilistic models. Because event and null environments were separated into two categories (i.e., 0 and 1), a binary response variable (Y), and therefore a logistic regression, was used to develop the probabilistic models.

2.5. Development of the Statistical Procedure

A successful severe weather statistical procedure (SWSP) was finalized after multiple iterations of different techniques to find the best statistical method for each step in the SWSP, when regarding the research problem, research goal, and compatibility with other methods in the procedure. The SWSP used a combination of statistical techniques while abiding by Occam's razor, which is the theory that a simpler approach minimizes the assumptions that need to be made and therefore is the better approach. The main steps in the procedure include a correlation matrix and dendrogram, clusters based on predictor versus predictor correlations, predictor selection based on the smallest deviance with Y, and logistic regression on the chosen predictors to produce a probabilistic equation with coefficients (weights) for each predictor. Other steps included model verification and evaluation to ensure a robust statistical model. The SWSP was written in R and some of the clustering code was adapted from Alfaro-Cordoba et al. (2017). R is a language and environment for statistical computing and graphics, and has numerous packages that can easily handle large datasets efficiently (R Core Team 2017). The procedure, listed out in the steps below, should be followed carefully and in the exact order stated. More details on specific R packages and functions can be found in the R documentation (R Core Team 2017). The ten-step SWSP is as follows:

Step 1) Choose and label cases:

Label event cases as Y=1 and null cases as Y=0; the binary logistic regression uses a binary response variable (Y). Each case is an individual row in the dataset. For this study, an event case was a HSLC convective environment where a severe weather hazard was officially reported, and a null case was a HSLC convective environment where a severe weather hazard was warned on but there was no official report of occurrence. As noted, the severe weather

hazards in this study were tornado and severe wind speed. This dataset, consisting of all events and nulls, is referred to as the "all" dataset. Divide the "all" dataset into tornado events and tornado nulls (i.e., "tornado" cases), and severe wind speed events and severe thunderstorm nulls (i.e., "severe wind speed" cases). Other test datasets were also created including "EF2+ tornadoes" dataset, which was requested by the NWS collaborators.

Step 2) Choose predictor list and fill in predictor values:

To fill in the predictor values, each dataset included the maximum value, minimum value, 3-hour difference maximum value, and 3-hour difference minimum value, as discussed in the previous section. These values were evaluated separately (e.g., maximum values of "tornado" cases were used as a separate training dataset) and then combined (i.e., maximum and minimum were combined into one dataset, and 3-hour difference maximum and 3-hour difference minimum were combined into another dataset) to examine if this would create better performing probabilistic models. Each predictor is an individual column in the dataset, along with a column for Y. The predictor values were calculated over a specific region (e.g., approximately 162x162 km^2 in Chapter II datasets) within a specific timespan (e.g., within 1.5 hours of the case time in the Chapter II datasets), based on the resolution of the gridded (NWP) model output. Be specific on the temporal and spatial spacings of the predictor values as part of their definitions; this will determine the temporal and spatial resolution of the input data when the probabilistic model is used operationally.

Step 3) Select test cases:

Randomly choose cases to pull out of the dataset for use later to validate the model; this is the testing dataset. The rest of the cases (not including the testing dataset) will be used as the

training dataset. The entire dataset is used for the first validation to make sure the model is working well. If the model cannot predict the training dataset, then a separate testing dataset would not be necessary. Separating the dataset by randomly choosing cases instead of separating the dataset by years is preferred because this limits bias when training a model. A case randomizer code was built into the procedure. A general rule of thumb is to remove 10% of the total cases from the dataset for a separate testing dataset, which would be around 200 cases for the "all" HSLC dataset. When testing, it is a good idea to have the same number of event cases as null cases, so 100 events and 100 nulls were randomly removed from the dataset and set aside as the testing dataset. The remaining 90% of the dataset was used as the training dataset.

To make sure the model validation is reliable, the testing dataset can be randomly selected multiple times; therefore, 200 random cases were tested in Test #1 and then another 200 random cases were tested in Test #2. When this was done with the "all" dataset, the results from both tests were consistent and did not change significantly, so it was decided that 200 testing cases was a sufficient sample size.

Step 4) Delete predictors (i.e., columns) with missing data (i.e., NAs):

This may only occur with calculated predictors during post-processing of the gridded (NWP) model output. If a predictor in the dataset contains missing values, this predictor will always have missing values due to the way the predictor is calculated (e.g., effective CAPE). These predictors should be removed because they, by default, do not provide a robust dataset due to being undefined variables at times. This step is necessary because most procedural steps will not work with missing values and it is also detrimental to include predictors in the final probabilistic models that do not have values at all gridded (NWP) model output times; therefore, the probability would not be calculated at that output time. If a predictor value is missing in only

a small number of cases, these cases can be deleted instead of deleting the entire predictor. This should not be done with predictors that always have missing values due to the way the predictor is calculated. The case deletions may theoretically weaken the relationships between the predictors due to decreasing the sample size. Note this method was tried, and deleting cases (i.e., rows) that contained NAs did not appear to change the resulting models from our dataset; these predictors were not chosen by the models when included in the dataset.

Step 5) Run clustering technique:

A correlation matrix with a dendrogram is created (e.g., Figure 2.1) to depict the relationship between predictors. The correlation matrix shows the Pearson's correlation coefficient (r) of each predictor vs. predictor combination for the entire list of predictors. The value of r ranges from -1.0 to 1.0. A correlation coefficient of 1 (or -1) indicates a perfect correlation (i.e., both predictors have a perfectly linear relationship), and a correlation of 0 indicates no relationship. In general, depending on the dataset, a r between 0.4 and 0.59 is considered a moderate correlation (linear relationship), and a r of 0.6 and higher is considered a strong linear relationship. The absolute value of r is used for clustering because the sign of r only indicates the direction of the relationship, which does not affect the closeness of the linear relationship between two predictors. If the absolute value of r is not used, the dendrogram will be arranged differently, because it considers the sign of the correlation. The absolute value needs to be used so the predictors are clustered solely on the closeness (absolute value of r) of the relationship between predictors.

The branches of the dendrogram indicate the closeness of the relationship between the predictors and provides height values for these branches. The user may choose where to "cut" the height of the dendrogram, which determines the number of clusters (i.e., groups of predictors).

These clusters group predictors that are highly correlated among each other, which are predictor-predictor correlations of 0.4 and higher (see Step 7 for more information). Note that the order of the predictors (columns) in the dataframe (dataset) is not relevant; the hcluster (clustering) code will find the best clusters based on absolute value of the correlation. The user should choose the number of clusters based on the predictor list and calculate the corresponding height; this allows the user to be cautious and not have clusters cut from different heights. The dendrogram is cut at a specific height, which is chosen by evaluating the correlation matrix and dendrogram; specifically, for the training datasets used in this research, 20-60 clusters were chosen for the first cut based on the predictor list. (The combined maximum value and minimum value datasets used the higher number of clusters.) When looking at the dendrogram (branches), a higher height chooses less clusters, which leads to less chosen predictors (see following steps). To allow for a substantial number of predictors to be chosen (to later be evaluated within statistical models), the lowest height possible should be chosen. However, a height too low would cause highly correlated predictors to be grouped into more than one cluster, which leads to more than one chosen predictor from a highly correlated group. This correlation clustering technique is similar to Alfaro-Cordoba (2017).

Step 6) Choose one predictor from each cluster:

Use the residual deviance (i.e., log-likelihood of full model) to determine the best fit, which is included in the summary of the model. The likelihood function ($\mathcal{L}$) is the maximum-likelihood estimate of the model. This method was chosen because it mathematically has the best results for a logistic regression model, which has a response variable (Y) that is not normal (i.e., it is binary).

The deviance is:

$$deviance = -2\log(\mathcal{L})$$

The deviance represents how much unexplained variation there is in the model, so the smallest deviance indicates the best model. Since each full model will include only one predictor (the predictor being compared) with the same dataset (i.e., same null deviance), the residual deviance can be compared between models to determine the most skillful predictor. The R code loops through one cluster at a time and chooses the predictor in the cluster with the smallest deviance with Y (e.g., Table 2.3). To do this, one at a time, a probabilistic model containing only one predictor in the cluster is calculated, the residual deviance is recorded, and then the next probabilistic model with a different predictor in the cluster is calculated and recorded. The lowest residual deviance is found out of all the probabilistic models calculated from that cluster, and that predictor is chosen. Therefore, one predictor is chosen from each cluster, and the code outputs the names of the chosen predictors. For analysis purposes, it may be useful to view which predictors are within each cluster.

Step 7) Double-check the correlations between predictors with matrix/dendrogram:

This double-checking step is important because of the nature of meteorological data, which is highly correlated, and it can be difficult to separate the predictors into precise groups (clusters). Due to the complicated correlations of atmospheric variables (i.e., many variables are linearly related), it may be necessary to run Steps 5 and 6 again to get rid of all highly correlated predictors left in the predictor list. Simply choosing less clusters at the start will not eliminate this problem because of the complicated relationships between multiple predictors (i.e., two predictors that are highly correlated with each other may not be in the same “branch” of the dendrogram where it is cut). Each time the dendrogram height is cut, the order of the variables

shifts in the correlation matrix, which allows the user to sift through the highly correlated predictor combinations. Any remaining highly correlated predictors should be removed using the smallest deviance with Y technique (Step 6), so both Steps 5 and 6 are repeated together. Step 7 should remove one predictor at a time, and may need to be repeated multiple times until all high correlations are gone.

The only user choice in this procedure is where to cut the height of the dendrogram, so this step will determine if there are still highly correlated predictors in the set of chosen predictors. Be wary of any predictor-predictor correlation above 0.4 and do not move on to the next step with any correlation above 0.5; this is done to minimize multicollinearity and limit the number of chosen predictors for the logistic regression step (Step 8). The 0.4 cutoff worked better with the training datasets in Chapter II, which had over 250 predictors. These correlation thresholds were determined by dataset evaluation (e.g., Tables 2.4-2.6), model evaluation (e.g., Chapter II probabilistic models that included predictor-predictor correlations above 0.4 contained predictors with p-values above 0.10), and a recommendation by Alfaro-Cordoba (2017). The final list of predictors is referred to as the "chosen predictors."

Step 8) Fit a logistic regression model with the chosen predictors:

With a binary logistic regression, the response variable (Y) assumes a binomial distribution and the link function is the logit transformation (Figure 2.2). This method was chosen due to the binary distribution of the cases, either an event case (Y=1) or a null case (Y=0), which gave a binary distribution of Y (specifically a Bernoulli distribution). The logistic regression produces a probabilistic equation, which will give a probability forecast of the severe weather hazard. See section 2.2 for more information on logistic regression and why this step (Step 8) was included in the SWSP.

A logistic regression models the logit-transformed probability as a linear relationship with the predictors. The logistic regression of Y on the predictors (x's) estimates the parameter values for the coefficients (β's) using the maximum-likelihood method. The $logit(\pi_i)$ is the logistic function of the probabilistic model, which is the natural log of the odds that Y=1:

$$logit(\pi_i) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_i$$

for $0 \leq p \leq 1$, where p is the probability of an event occurring (i.e., the probability that Y=1).

This step gives a final probabilistic equation (below), including the names of the chosen predictors (that were plugged into the logistic regression) and the corresponding β (coefficients) for each predictor. A coefficient is the weight assigned to the predictor within the probabilistic equation; it is a maximum-likelihood estimator. The following equation shows how the chosen predictors, X_i with i = 1, 2, …, and the corresponding coefficient (i.e., weight) of the predictor, β_n with n = 0, 1, 2, …, are plugged into the probabilistic model, π_i:

$$\pi_i = \Pr(Y_i = 1 \mid X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_i)}$$

where β_o is the intercept, X_1 is the value of predictor 1, β_1 is the coefficient of predictor 1, X_2 is the value of predictor 2, β_2 is the coefficient of predictor 2, and continues for the number of predictors in the probabilistic model. This equation produces a probability between 0 and 1, where values closer to 1 represent an event and values closer to 0 represent a null. The probability threshold may not be exactly 0.5 and it is determined by the optimal threshold code for each model (see Step 10). All predictors may be kept in the model since best fit procedures were already used in previous steps.

The logistic regression identifies linear aspects of predictability with statistical significance, where significance is defined as a p-value less than 0.05 (i.e., 95% confidence

level). Looking at the significance (p-value) for each predictor, the user can determine if predictors should be dropped from the model; this could indicate predictors that do not have a strong relationship with Y. If the model has "too many" explanatory variables (predictors), the model could falsely give a significant p-value (i.e., Type I error). To decide on the best model, drop the predictor in question and re-compute the probabilistic model. Then compare the Akaike Information Criterion (AIC) between the models, i.e., the model with all predictors versus the re-computed model(s); the model with the lowest AIC is the best model. The AIC is:

$$AIC = -2\log(\mathcal{L}) + 2p$$

where p is the number of estimated parameters (predictors, the intercept, and error term) in the model and $\boldsymbol{\mathcal{L}}$ is the maximum-likelihood for the model. Thus, the model with the lowest AIC is the best model. The 2p term is the penalty, which increases with an increase in the number of predictors in the model; this this helps determine the simplest most skillful model. Therefore, one model could have a lower deviance than another model, but have a higher AIC due to the penalty for more predictors. The AIC should only be compared between models that were produced with the same training dataset (same cases), because the likelihood will change with a change in datapoints (i.e., cases). The AIC is not interpretable on its own. Comparing AIC values will determine the number of predictors in the model by deciding which model is the most skillful; therefore, the number of predictors in skillful probabilistic models can vary and it is dependent on the predictive information in the dataset. The AIC penalizes for more complicated models (i.e., the number of predictors in the model), so it prevents the user from choosing a model with extraneous predictors. It is possible for a predictor to have a p-value greater than 0.05 in the model and be kept in the model, if the AIC comparison shows that the predictor increases the

model skill. Once the best model is decided, record the predictor names and the corresponding coefficients that define the final probabilistic equation.

It is important to note that the sign of the coefficients in the probabilistic model may seem counterintuitive because the predictor is increasing/decreasing at the same time the other predictors in the model are increasing/decreasing. For more information on the predictors within the probabilistic model, compute how the odds of Y changes as a function of the predictors by calculating the odds-ratios. For example, if the odds-ratio corresponding to a predictor is 1.2, then increasing the predictor value by one unit will increase the odds of Y=1 by 0.2, assuming all the other predictors in the model are held constant. In other words, to interpret an individual coefficient, all the other predictors in the model must be held constant. The 95% confidence intervals for the odds-ratios can be calculated to show the uncertainty in the estimate; i.e., exp(confint(model)). Note that "model" in the code refers to the probabilistic model containing a single predictor, the predictor being evaluated. Because a predictor is likely to increase/decrease at the same time of an increase/decrease of another predictor in the model, these odds-ratios may not give useful information. Therefore, it is best to examine the probabilistic model as one holistic system (Table 2.13).

Step 9) Use testing datasets to test the final probabilistic model

The first dataset used to validate the model is the testing dataset that was put aside in Step 3, which gives a less biased evaluation (compared to the training dataset). The testing dataset is comprised of different cases than the cases in the training dataset, so the reliability of the model can be evaluated. The second dataset used to evaluate the model is the training dataset used to create the model. This will test how well the probabilistic model output matches the known results (i.e., analysis input data). For forecast verification, the third dataset is the gridded (NWP)

model output at the 1-hour forecast (i.e., use forecast data instead of analysis data); different forecast times can also be used to test lead times. We can also test the model using a dataset consisting of values from a different gridded (NWP) model to test the versatility of the statistical model. For brevity, Chapter II focused on the first two datasets mentioned here. The first three datasets were used in subsequent chapters.

Step 10) Validate and evaluate the final probabilistic model:

First, check to make sure the probabilistic model does not violate assumptions. The logistic regression assumptions include: 1) the outcome is a binary variable (e.g., 0 or 1), 2) there is a linear relationship between the logit of the outcome and each predictor (i.e., logit(p) = log(p/(1-p)) where p is the probability of an event occurring), 3) there are no extreme values (outliers) in the continuous predictors (i.e., the absolute standardized residuals are below three), and 4) there are no high intercorrelations among the predictors (i.e., multicollinearity) (Kassambara 2017). Multicollinearity was resolved in Steps 5 and 6, and then double checked in Step 7; it can be examined by calculating the variance inflation factor (VIF), which estimates the severity of multicollinearity and should not exceed five. Temporal and spatial correlation errors can be checked manually by comparing the dates and times of the cases. Temporal correlation and spatial correlation are related in these training datasets due to the nature of the samples (i.e., predictor values in a specific spatial and temporal domain), so correcting for temporal correlation error also minimized spatial correlation error. (Since we are not using time series data and the predictors are not all linearly related, it would not be appropriate to check temporal/spatial correlation with the Auto Correlation Function or PCA methods.)

Test the final probabilistic models by inputting the predictor values from the testing dataset (part of Step 9). Regarding the predictive skill of a statistical model, a confusion matrix is

a useful evaluation method to present to potential users, such as National Weather Service (NWS) forecasters. This technique is similar to skill scores that are currently used by the NWS (e.g., Heidke Skill Score; NOAA NWS 2018), and it explicitly states the false alarm rate and number of missed events predicted by the statistical model (in the dataset given). A confusion matrix illustrates true positives (correctly predicted events), true negatives (correctly predicted nulls), false positives (false alarms), and false negatives (missed events) with varying probability thresholds for each model (Figure 1.1). Before calculating the confusion matrix, determine the optimal cutoff for the probabilistic model. The optimal cutoff is the probability threshold between labeling the case as an event or as a null (i.e., binary classifier). The optimal cutoff (probability threshold) is determined by the lowest misclassification error (percentage of missed cases) possible for the model. For example, if the model had an optimal cutoff of 0.50, to get the highest number of correctly predicted cases, a probability above 0.50 predicts an event to occur and a probability below 0.50 predicts a null. Figure 1.1 illustrates this with a general example of a confusion matrix.

The receiver operating characteristic curve (ROC curve) illustrates the skill (robustness) of a probabilistic model by plotting the true positive rate (TPR) versus the false positive rate (FPR). The TPR is also referred to as the probability of detection (POD), and the FPR is also referred to as the false alarm rate (FAR); therefore, the ROC curve visualizes the POD versus FAR of a model (Figure 1.2). The area under the ROC curve (AUROC), ranging from 0 to 1, is used to determine skill in a statistical model with a value of 1 representing a perfectly skilled model (i.e., no error) and 0 representing a model with no skill. A model earning a value of 0.5 is considered to have skill equivalent to random chance; therefore, a model will be considered

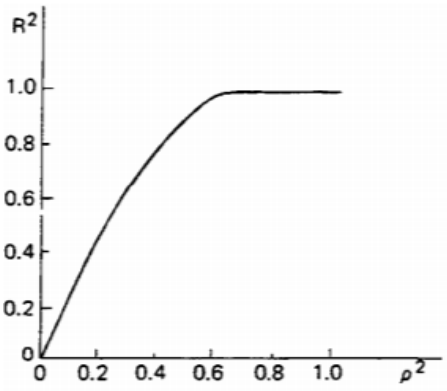
skillful if the AUROC curve is greater than 0.5. An AUROC is 1 when POD is 100% and FAR is 0%.

The misclassification error and AUROC were used as two main model evaluation tools for the probabilistic models. The model with the lowest misclassification error was the model that correctly predicted the most cases (events and nulls). The optimal cutoff (probability threshold) was determined by the lowest misclassification error for the model. The model with the highest AUROC was the most skillful model, which was based on the POD versus FAR of the model. AUROC represents overall model skill. Lowering the optimal cutoff of a model also lowers both POD and FAR (Figure 1.1), which can also be seen when following down along the ROC curve (Figure 1.2); therefore, the optimal cutoff is the point on the ROC curve that maximizes the number of correctly predicted cases. Other validation techniques were used, but for brevity, the confusion matrix and AUROC values were reported. If desired, cost functions can be used to weigh missed events more/less (greater/smaller penalty) than false alarms, which allows for a harsher technique based on the user's concerns; for example, the NWS forecasters could weigh missed events more than missed nulls (false alarms).

The predictors in the model can be centered and scaled, so the coefficients (β) can be compared. The coefficients should not be compared unless the predictors were centered and scaled before running the logistic regression. After centering and scaling, the predictor with the largest coefficient is the most influential predictor in the model (i.e., contributes the most to the prediction). This centered and scaled coefficient represents the contribution of the predictor, given that all the other predictors are already in the model, so this is showing the extra contribution the predictor gives to the model. This means the predictor could be contributing more, but part of the contribution was accounted for by the other predictors in the model (same

predictive information). This explains why a predictor's coefficient will decrease when more predictors are added to the model. A positive coefficient indicates the predictor increases the probability (of a severe weather hazard) and a negative coefficient indicates the predictor decreases the probability, assuming all other predictors are held constant. Adding up the absolute value of the coefficients of the predictors (β's) reveal how much of the outcome (Y) is explained by the model. For example, if the β's of the predictors (not including intercept β_0) add up to 0.36, the predictors in the model combined contribute 36% of the information needed to fully predict Y (i.e., explained variation in Y).

Another way to assess the explained variation in Y by the model is by calculating the McFadden's Pseudo-R^2, a measure of goodness of fit. The coefficient of determination, R^2, used in ordinary linear regression does not work for logistic regression, which is why a "pseudo" R^2 is calculated. McFadden's Pseudo-R^2 (ρR^2) is defined as 1 minus the ratio of the log-likelihood of full model to the log-likelihood of null model; therefore, it can be calculated using the residual deviance (log-likelihood of full model) and null deviance (log-likelihood of null model) given in the model summary. This gives an estimate of the total explained variation in Y by the model. As shown in the graph below (from Domencich and McFadden 1996), ρR^2 values tend to be substantially lower in comparison to R^2, so a ρ R^2 value between 0.2 and 0.4 is considered to represent an "excellent model fit":



The 95% confidence interval of the model's prediction (probability) can be calculated to visualize the uncertainty in the probability given by the model. This uncertainty is due to the uncertainties of each predictor. For example, dropping a predictor from the model can lower the uncertainty of the model; less predictors usually equate to less uncertainty in the prediction. This gives the overall model uncertainty (for all predicted probabilities). See Table 2.10 for a way to provide a more detailed analysis on the predicted probabilities.

2.6. Results & Discussion

A geographical visualization of all cases included in the "all" dataset (2063 cases) is shown in Figure 2.3. Most null cases were near the coastline and/or at lower elevations, which could be due to observation bias (i.e., these are more populated areas). Geographical maps showing the locations of all cases in the "tornado" dataset (865 cases) are shown in Figure 2.4. Majority of cases in the dataset occurred during meteorological winter and spring seasons, which is also true for the "severe wind speed" dataset (1198 cases). Figure 2.5 shows the locations of all cases in the "severe wind speed" dataset, along with the time of the storm report or null case. In both the "tornado" and "severe wind speed" datasets, most cases occurred in the evening or overnight hours. A quick overview of the results produced from each step of the SWSP, including the probabilistic models, using the dataset described in sections 2.3 and 2.4 (i.e., Sherburn's NARR 5x5 dataset), are shown and discussed in this section.

Note that all probabilistic models in this section were evaluated at their optimal cutoff (i.e., probability threshold with lowest misclassification error) to show the highest possible skill of the model, for a fair comparison. Therefore, the misclassification error and AUROC (overall model skill) are used to determine model skill.

2.6.1. Proof of Concept for Predictor Selection in SWSP

The Sherburn et al. (2016) study is a HSLC severe weather hazard prediction study in SEUS, similar to this research study, so their dataset was used to test the SWSP. The "all" dataset (i.e., all cases) including all predictors available was used in Sherburn et al. (2016), so this same dataset was used as the proof of concept of the SWSP, to be consistent with the evaluation completed by Sherburn et al. (2016). Therefore, the most skillful predictor list via maximum Heidke Skill Score (HSS) shown in Table 2.2 (Sherburn et al. 2016) can be reasonably compared to the most skillful predictor list via smallest deviance with Y (Table 2.3). More specifically, the smallest residual deviance was used to determine the best fit with Y (i.e., finding the maximum likelihood), which is Step 6 of the SWSP. In the Sherburn et al. (2016) study, the predictor list was split into two categories, ingredients and combination index; this was based on my concerns early on in their study that predictors made up of multiple variables (i.e., composite index) may, in some circumstances, artificially show more skill due to high correlations between the individual variables within the composite index. The top ten "ingredient" predictors, in order, were: UESHR, U2SHR, HPBLSFC, FGEN725, EFFSHR, FGEN750, W10M, V10M, FGEN700, LR03. The top ten "combination index" predictors, in order, were: SHERBE, SHERB35, TKE950, TKE900, TKE925, TKE875, TEVV7, TEVV45K, TEVV3K, TEVV5. Further information on these predictors is given in Table 2.1. The top 9 "ingredient" and top 10 "combination index" predictors were within the top 34 predictors out of 253 (not including MOSH and MOSHE for an equivalent comparison), shown in Table 2.3. Because this same dataset was used to develop the MOSH and MOSHE, it is no surprise that these two predictors were the top two most skillful predictors in Table 2.3. These results show that the smallest deviance with Y method sufficiently identifies the most skillful predictors, and it also highlights

the need for sorting through the highly correlated predictors to find the top predictors that do not repeat information (i.e., more independent information) in the chosen skillful predictors (i.e., Step 5 of the SWSP).

Because the smallest deviance with Y method directly measures the relationship of each predictor with Y, it is a more direct way of finding skillful predictors. This relationship between the predictor and Y should be relatively unchanged if additional HSLC cases in SEUS were to be evaluated (assuming the same distribution of tornado and severe wind cases if using the "all" dataset). The correlation matrix, dendrogram, and clustering step (Step 5) sorts through the most skillful predictors to provide a final predictor list that does not include highly correlated predictors (e.g., Tables 2.4-2.6), which ensures that most of the skill available in the variable list is kept in the final list of skillful predictors without including predictors that share the same predictive information (i.e., highly correlated). For a more efficient procedure, Steps 5-7 are in the same section of R code, which is explained further in the Step 7 instructions. The combination of the smallest deviance with Y and the correlation/dendrogram clustering is an objective way of selecting skillful predictors for the probabilistic model.

Because predictor selection techniques that are typically used with logistic regression rely on significant p-values for choosing skillful predictors, those techniques were not utilized in the SWSP. Also, some predictor selection techniques (e.g., backward stepwise) do not necessarily chosen unique predictors (Marzban et al. 1999). During development of the SWSP, backwards selection, forwards selection, and LASSO were tried; all techniques had issues. For example, using LASSO on the chosen predictors resulted in LASSO choosing just the first two predictors in the list for the final model, because all chosen predictors were skillful, and LASSO penalizes based on the total number of predictors in the final model. It was found that the LASSO predictor

selection was too aggressive and did not provide the most skillful models. When producing statistical models, a common misunderstanding is that a lower p-value equates to a more skillful predictor, which may not be true. After numerous trials, it was found that p-values for predictors changed substantially based on the predictors in the model, which is related to predictors having high correlations among each other. After selecting skillful predictors from Steps 5-7 of the SWSP, an additional predictor selection technique was not needed. Therefore, after the skillful predictors are selected from Steps 5-7, they are inputted into a logistic regression. Then, to determine if any additional predictors should be dropped from the model, the AIC of models were compared, which is discussed more in Step 10.

2.6.2. Utilizing Maximum Value "All" Training Dataset Including Composite Parameters

To provide a proof of concept for the SWSP as a whole system, the maximum value "all" dataset including all variables was used again. For comparison to Sherburn et al. (2016), the entire variable list (Table 2.1) was evaluated, even though some of the variables are not reasonable for a predictive equation and some variables are severe weather composite parameters (e.g., STP). Variables that were removed from the predictor list due to NAs were MLCIN, MUEL, SBLFC, CLBPRS, EFFNCAPE, EFFBOT, EFFTOP, NHGTCLB, PRATE; see Step 4 and Table 2.1 for more details.

Based on the dendrogram including all variables in Table 2.1 (except for the 9 variables removed due to NAs), 15 clusters were chosen for the first cut of the dendrogram in Step 5, which was a height of 4.275546. Step 6 revealed the top 15 predictors as: SPFH1815, PWAT, TEDF02, DIVG950, MOSHE, HGTTRP, FGEN800, U2SHR, DIVG250, CAPE05, SHERBE, SBLI800, LR03, TEVV7, U10M (see Table 2.1). Step 7 was completed three times until all

correlations were below 0.4. When completed, the highest correlation (0.3371) was between U2SHR and MOSHE. Detailed graphics of the final dendrogram and correlation matrix are shown in Figure 2.6 (showing the direction of the correlation for additional information, but that was not used in the SWSP). Step 8 gives the final probabilistic equation, including the names of the significant predictors and the corresponding β (coefficients) for each predictor.

The predictors in the final model and their coefficients (i.e., Estimate) are shown in the R output (see Table 2.14):

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.809e-01  4.691e-01   1.239 0.215514
HGTTRP      -1.363e-04  2.914e-05  -4.678 2.90e-06 ***
U10M         9.166e-02  1.721e-02   5.327 9.96e-08 ***
DIVG950     -3.536e+03  1.954e+03  -1.810 0.070304 .
MOSHE        2.894e-01  3.473e-02   8.332  < 2e-16 ***
FGEN800      1.446e-01  3.079e-02   4.698 2.63e-06 ***
U2SHR        3.666e-02  1.087e-02   3.372 0.000746 ***
DIVG250      5.185e+03  7.897e+02   6.566 5.18e-11 ***
SBLI800     -1.226e-01  1.864e-02  -6.575 4.85e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2690.6  on 2062  degrees of freedom
Residual deviance: 2085.2  on 2054  degrees of freedom
AIC: 2103.2
```

The chosen predictors make physical sense and appear to represent a dynamic system at the mesoscale that would support and/or describe a HSLC severe storm environment in the SEUS, including convergence near the surface (DIVG950), divergence aloft near the jet-level (DIVG250), tropopause height (HGTTRP), low-level wind shear (U2SHR), surface winds in the west-east direction (U10M), frontogenesis (FGEN800), lifted index within the planetary boundary layer (SBLI800), and MOSHE (indicating chance of severe weather hazards within HSLC convective environments). For reference, the planetary boundary layer is discussed in Figure 2.7.

When validating the final model against the testing dataset (of 200 random cases), the AUROC was 0.8126 and the misclassification error was 24.0%; there were 36 false alarms and 12 missed events. These values were similar when evaluating other sets of 200 random cases, to test reliability. When validating against the entire dataset (of 2063 cases), the AUROC was 0.8083 and the misclassification error was 24.0%; there were 299 false alarms (439 true nulls) and 196 missed events (1129 true events). This was a false alarm rate of 40.5% (at the optimal cutoff). The POD was 85.2% (see sections 2.7 and 2.8 for discussion). The final model had an optimal cutoff of 0.510; therefore, to get the best results, a probability above 0.510 was predicted as an event and a probability below this cutoff was predicted as a null. The optimal cutoff was determined by the lowest misclassification error possible. The 196 missed events may seem like a large number (17% of events), but some of these cases were NWS missed events and/or were near the 0.510 cutoff value, which is one reason why probabilities are more useful than a yes (event) or no (null) answer. Out of the 196 missed events, 145 cases had a MOSHE of less than 1.0, which indicated no events expected (based on the MOSHE threshold recommended in Sherburn et al. 2016). The MOSHE composite parameter includes two shear predictors, S15MG and EFFSHR, as well as two other predictors, LR03 and MAXTEVV. The only shear predictor in the probabilistic model, U2SHR, ranged from 2.6 to 27 m/s in the missed events, 0 to 34 m/s in the missed nulls, and 0 to 35 m/s in the entire dataset. Bulk wind shear (e.g., EFFSHR) is a good indicator of potential severe weather hazards, but it can contribute to false alarms (personal communications with NWS forecasters). Shear predictors are useful in highlighting areas of potential severe weather (e.g., what the MOSHE is utilized for), but they may not be as useful in predicting the probability of a severe weather hazard (section 3.6.4); however, this may not be true for low-level shear predictors (e.g., U2SHR) so more testing is needed.

Step 8 in the SWSP includes investigating any predictor left in the model that has a p-value greater than 0.05, such as DIVG950 in this previous model. Dropping the DIVG950 term raised the AIC slightly to 2104.5, a less skillful model, which added 9 more false alarms and 3 more missed events. But, this small increase in skill may not justify keeping this addition predictor since additional predictors bring more uncertainty into the model. The AIC is more reliable than the p-value when deciding to remove a predictor in a logistic regression, so the DIVG950 term was kept in the model; however, the AIC was only slightly different, so it would be justifiable to remove the predictor.

During model validation (Step 10), the predictors were centered and scaled to reveal the contribution of each predictor in the model. The predictors combined contributed 35.8% of the information needed to fully predict Y (i.e., explained variation in Y). This means that other information (including other predictors) needed for the prediction were not included, which may include predictors that were not in the training dataset. The list of predictors in order of largest to smallest contribution to the prediction were: SBLI800, DIVG250, U2SHR, MOSHE, U10M, FGEN800, HGTTRP, DIVG950 (see Table 2.1). The scaled coefficient represented the contribution of the predictor, given that all the other predictors were already in the model, so this showed the extra contribution the predictor gave to the model. This means the predictor could be contributing more to the prediction, but part of the contribution was accounted for by the other predictors in the model (same predictive information). For example, MOSHE explained 6.3% of the variation in Y in the model and 14.6% of the variation in Y on its own. SBLI800 explained 9.6% of the variation in Y in the model and 12.0% of the variation in Y on its own. The MOSHE had a closer relationship with Y (smaller deviance with Y) than SBLI800, which helps explicate why it explained more variation in Y on its own. However, the SBLI800 brought more “original”

predictive information (i.e., more independent information) to the model than MOSHE, which is why it explained more variation in Y in the model (i.e., SBLI800 contributed more to the model).

The 95% confidence interval of the model's prediction (probability) revealed an uncertainty of +/- 0.080; there is a 95% confidence that the probability (%) given by the model is +/- 8.0% of the "actual" probability. For example, if the predicted probability was 0.50, the 95% confidence interval is 0.42 to 0.58, which means the probability is between 42%-58% when accounting for uncertainty at the 95% confidence level ($\alpha = 0.05$). Note that this uncertainty is due to the uncertainties of each predictor. For example, dropping a predictor from the model can lower the uncertainty of the model; less predictors usually equate to less uncertainty in the prediction. Therefore, the "actual" probability refers to the prediction (probability) if there were no uncertainties in the predictors. The model only provides 35.8% of the prediction, so the literal "actual probability" can only be calculated if the model provided 100% of the prediction; however, this would be a perfect model, which does not exist with any severe weather forecasting model.

Note that slight variations in the predictors can provide a model with the same overall skill (same AIC) but change the confusion matrix. For example, if the DIVG950 term was not included in the previous model, the AIC was 2104.5. If FGEN800 was switched with FGEN725, the probabilistic model gives an AIC of 2104.4.

This is shown in the R output (see Table 2.15):

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.826e-01  4.673e-01   0.819 0.412966
HGTTRP      -1.353e-04  2.902e-05  -4.662 3.12e-06 ***
MOSHE        2.939e-01  3.457e-02   8.502  < 2e-16 ***
FGEN725      2.054e-01  4.179e-02   4.915 8.89e-07 ***
SBLI800     -1.076e-01  1.799e-02  -5.979 2.24e-09 ***
U2SHR        3.769e-02  1.080e-02   3.490 0.000483 ***
U10M         1.003e-01  1.695e-02   5.917 3.28e-09 ***
DIVG250      4.988e+03  7.840e+02   6.362 1.99e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2690.6  on 2062  degrees of freedom
Residual deviance: 2088.4  on 2055  degrees of freedom
AIC: 2104.4
```

When validating against the entire dataset (of 2063 cases), the AUROC was 0.8087 and the misclassification error was 24.9% (compared to 24.6% with FGEN800); there were 294 false alarms (444 true nulls) and 219 missed events (1106 true events). The final model had an optimal cutoff of 0.530. This model predicted 14 fewer false alarms and 20 more missed events than the model with FGEN800 included. FGEN800 and FGEN725 have a correlation of 0.72, which is considered highly correlated. This gives an example of how highly correlated predictors provide similar skill in probabilistic models. However, FGEN725 was not highly correlated with the other predictors in the model and it had a small deviance with Y also, which allowed the model skill to remain relatively unchanged. It is not advised to switch a predictor in the model with another highly correlated predictor without checking all predictor versus predictor correlations within the model. Switching a predictor, like the FGEN800 and FGEN725 switch, will most likely raise the misclassification error as shown in the previous example. This is because the SWSP was developed to choose the most skillful predictors with respect to the response variable (Y) using the smallest deviance with Y.

Another consideration is deciding if it is better for the probabilistic model to predict the lowest number of false alarms or the lowest number of missed events. The misclassification error can only improve if the model skill improves, so changing the optimal cutoff value will change the ratio of false alarms to missed events in the confusion matrix. If the model skill does not change, lowering the false alarms will increase the missed events, and vice versa. This can be visualized using the example in Figure 1.1.

2.6.3. Evaluating Severe Weather Composite Parameters for Probabilistic Prediction

The Sherburn et al. (2016) study produced two HSLC severe weather composite parameters, MOSH and MOSHE, using the same training dataset ("all" dataset) so these were evaluated using logistic regression to assess the probabilistic prediction capabilities. Sherburn et al. (2014) produced two HSLC severe weather composite parameters, SHERB and SHERBE, which did not perform as well as the MOSH and MOSHE (i.e., modified SHERB) in various evaluations, so those results were not shown. As hinted in the results from section 2.6.2, MOSHE explained 14.6% of the variation in Y on its own (i.e., a probabilistic model with MOSHE as the only predictor). When validating against the entire dataset (of 2063 cases), the AUROC was 0.7487 and the misclassification error was 29.0%; there were 254 false alarms (484 true nulls) and 345 missed events (980 true events). When the predictors inside the MOSHE (i.e., LR03, S15MG, EFFSHR, and MAXTEVV) were individually plugged into the model as separate predictors, the predictors explained 33.1% of the variation in Y, the AUROC was 0.7555 and the misclassification error was 28.8%; there were 417 false alarms (321 true nulls) and 178 missed events (1147 true events). The comparison gives two main points: MOSHE may perform better as a probabilistic model when the individual predictors are weighted based on maximum-

likelihood, and/or a composite parameter probabilistic model does not perform as well as a probabilistic model with its individual terms as predictors. Figure 2.8 gives some insight into why the last point may be true by showing a visual representation of the MOSHE probabilistic model and a probabilistic model containing only EFFSHR (a component of MOSHE). It is also possible that Sherburn et al. (2016) weighted the predictors within the MOSHE to reduce the false alarms without taking into account the concurrent increase in missed events. (Note the only predictor-predictor correlation above 0.4 in MOSHE was MAXTEVV and EFFSHR with a correlation of 0.4009, so highly correlated predictors were not an important factor.)

MOSH explained 16.4% of the variation in Y (on its own). (Smallest deviance in Y and explained variation in Y are different, which was why there was a slight difference here, if comparing to Table 2.3.) When validating against the entire dataset (of 2063 cases), the AUROC was 0.7533 and the misclassification error was 27.1%; there were 321 false alarms (417 true nulls) and 238 missed events (1087 true events). When the predictors inside the MOSH (i.e., LR03, S15MG, and MAXTEVV) were individually plugged into the model as separate predictors, the AUROC was 0.7519 and the misclassification error was 28.9%; there were 347 false alarms (391 true nulls) and 249 missed events (1076 true events). The opposite was true with the MOSH model comparison compared to the MOSHE model comparison. However, the model with separated predictors explained more variation in Y compared to the MOSH model, which was also seen with the MOSHE model comparison (which makes intuitive sense).

All four probabilistic models above were less skillful compared to the probabilistic model produced by the SWSP. This indicates that the predictor selection methods used in the SWSP (i.e., the only difference between probabilistic model development) kept more predictive information in the probabilistic model compared to the HSS predictor selection method; adding

to the proof of concept for the SWSP predictor selection methods (section 2.6.1). Also, when comparing all four probabilistic models, it was discovered that EFFSHR did not add substantial skill to the model and contributed to more false alarms when added to the model. (Another qualitative assessment showed the model with EFFSHR had only 0.3% more explained variation in Y compared to the model with EFFSHR removed, but this should not be used an indicator of model performance.) It has been noted by some NWS forecasters that bulk wind shear terms tend to contribute to false alarms, so this may be related (personal communications with NWS).

To investigate the probabilistic prediction capabilities of other severe weather composite parameters, in general, the STP, SCP, VGP, and EHI were evaluated (see section 2.4 for more information on these composite parameters). These predictors were evaluated separately by producing a probabilistic model with just the one predictor being evaluated, for a fair comparison. STP explained 8.2% of the variation in Y. When validating against the entire dataset (of 2063 cases), the AUROC was 0.6428 and the misclassification error was 34.7%; there were 443 false alarms (295 true nulls) and 273 missed events (1052 true events). SCP explained 9.3% of the variation in Y. For SCP, the AUROC was 0.6332 and the misclassification error was 35.5%; there were 492 false alarms (246 true nulls) and 241 missed events (1084 true events). VGP explained 10.6% of the variation in Y. For VGP, the AUROC was 0.6421 and the misclassification error was 33.6%; there were 648 false alarms (90 true nulls) and 45 missed events (1280 true events). EHI explained 8.2% of the variation in Y. For EHI, the AUROC was 0.6155 and the misclassification error was 34.1%; there were 614 false alarms (124 true nulls) and 89 missed events (1236 true events).

The misclassification errors of these four severe weather composite parameters were relatively high, ranging from 33.6% to 35.5%. The AUROC of these probabilistic models ranged

from 0.6155 to 0.6428, which is considered low model skill; 0.5 is considered to have skill equivalent to random chance. For comparison, the probabilistic model produced by the SWSP, the AUROC was 0.8083 and the misclassification error was 24.0%; there were 299 false alarms (439 true nulls) and 196 missed events (1129 true events). Validation techniques, including the confusion matrix and AUROC, indicate currently used severe weather composite parameters (i.e., STP, SCP, VGP, and EHI) are not skillful in probabilistic prediction of HSLC convective environments, which shows the importance of designing probabilistic models for specific convective environments like HSLC. For example, the STP and SCP were developed using only supercellular environments, which also focused on HSHC convective environments. The MOSHE and MOSH severe weather composite parameters were specifically designed for HSLC convective environments and this is shown by the increased model skill compared to the other composite parameters designed for severe convective environments. The MOSHE and MOSH were developed using the "all" dataset but showed relatively the same model skill with an independent testing dataset (see Sherburn et al. 2016), so while this might provide bias, it would not explain the drastic difference in model skill seen in the comparisons with the other severe weather composite parameters.

Looking at the severe weather composite parameters mentioned here, they were all designed to indicate severity and/or distinction between severe and non-severe environments of some type. Some severe weather composite parameters highlight areas with favorable environments for severe weather hazards. They were not designed for predicting the interpretable probability of an individual severe weather hazard (e.g., tornado). These severe weather composite parameters were made up of components (predictors) that were already included in the predictor list, so they provide the same predictive information in the training dataset. This

also appears to make the correlation matrix and dendrogram slightly more complex (e.g., comparing Figure 2.1 with Figure 2.9) which causes multiple reiterations of Step 7. Another potential issue is some severe weather composite parameters can be undefined at times (e.g., the square root of CAPE is a component of VGP, which would be undefined when CAPE equals zero), which would need to be set to zero to avoid missing values (NAs). If a large proportion of the cases have a zero value for a predictor, this could decrease the skill of the predictor in the model.

Since the training dataset needs to be designed specifically to meet all the assumptions of the statistical methods used to produce a probabilistic model, it is best to start from scratch when developing the training dataset for probabilistic models. In the case of the MOSHE, the individual components (predictors) of the MOSHE performed better than the MOSHE as a single predictor in the probabilistic model. Since the MOSHE was highly correlated with two of its components, EFFSHR and MAXTEVV (both had over 0.55 correlation), it was clustered with these two predictors and chosen to represent the cluster over the other two predictors, which changed the outcome.

To test what happens when severe weather composite parameters are not included in the training dataset, they were removed. Based on the dendrogram, 30 clusters were chosen for the first cut of the dendrogram. After removing predictors with correlations above 0.4, the top (13) predictors (in order of smallest to largest deviance with Y) were: TKE925, UESHR, FGEN725, SBLI800, DIVG250, HGTTRP, MULFC, DIVG950, TMP1000, CAPE05, VVEL975, DIVG750, VVEL700 (see Table 2.1).

The predictors in the model are shown in the R output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.214e+00  6.398e+00   0.971 0.331493
HGTTRP      -1.548e-04  3.143e-05  -4.926 8.39e-07 ***
TMP1000     -1.963e-02  2.227e-02  -0.882 0.377928
VVEL975     -6.200e-01  4.205e-01  -1.475 0.140334
VVEL700      2.691e-01  1.846e-01   1.458 0.144932
DIVG950     -5.475e+03  2.134e+03  -2.566 0.010287 *
TKE925       4.527e-01  4.101e-02  11.038  < 2e-16 ***
DIVG750     -4.819e+02  2.472e+03  -0.195 0.845447
FGEN725      2.173e-01  4.508e-02   4.821 1.43e-06 ***
DIVG250      5.325e+03  8.544e+02   6.233 4.58e-10 ***
CAPE05       9.712e-05  9.177e-04   0.106 0.915718
SBLI800     -6.883e-02  2.017e-02  -3.413 0.000643 ***
MULFC       -1.131e-04  2.443e-05  -4.630 3.65e-06 ***
UESHR        7.235e-02  8.841e-03   8.184 2.75e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2690.6  on 2062  degrees of freedom
Residual deviance: 2036.6  on 2049  degrees of freedom
AIC: 2064.6
```

In this case, the predictors with p-values greater than 0.05 are the predictors with the largest deviance in Y (i.e., TMP1000, CAPE05, VVEL975, DIVG750, and VVEL700). When removing these predictors from the model, the model skill improved, shown by a lower AIC.

The predictors in the final model and their coefficients (Estimate) are shown in the R output (see Table 2.16):

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.389e-01  4.584e-01   1.394 0.163397
TKE925       4.387e-01  3.848e-02  11.401  < 2e-16 ***
UESHR        6.880e-02  7.905e-03   8.704  < 2e-16 ***
FGEN725      2.230e-01  4.381e-02   5.090 3.59e-07 ***
SBLI800     -6.776e-02  1.842e-02  -3.680 0.000234 ***
DIVG250      5.059e+03  8.281e+02   6.109 1.00e-09 ***
HGTTRP      -1.659e-04  2.955e-05  -5.615 1.97e-08 ***
MULFC       -1.135e-04  2.435e-05  -4.660 3.16e-06 ***
DIVG950     -4.769e+03  1.991e+03  -2.396 0.016584 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2690.6  on 2062  degrees of freedom
Residual deviance: 2040.9  on 2054  degrees of freedom
AIC: 2058.9
```

To test if this is the best model, DIVG950 was removed. The AIC increased to 2062.7, a decrease in model skill, so DIVG950 was kept in the model. The optimal cutoff of the model was 0.4588. The 95% confidence interval of the model’s prediction (probability) revealed an uncertainty of +/- 0.077. When validating against the entire dataset (of 2063 cases), the AUROC was 0.8175 and the misclassification error was 23.2%; there were 352 false alarms (386 true nulls) and 127 missed events (1198 true events). The predictors in the final model were the same or similar to the predictors in the most skillful model produced with the training dataset that included the severe weather composite parameters. For reference, that model had an AUROC was 0.8083 and the misclassification error was 24.0%; there were 299 false alarms (439 true nulls) and 196 missed events (1129 true events). Not only were the severe weather composite parameters not necessary, but the model skill improved slightly.

This gives evidence that severe weather composite parameters are not necessary and are not recommended for inclusion in the training datasets, if predictors related to the components of

the composite parameter are in the training dataset. Technically the TEVV terms were calculated using other predictors in the predictor list, so these terms and others like it may need to be removed from the predictor list for similar reasons to the removal of severe weather composite parameters. This should be tested in future work. (This was kept in mind during the creation of training datasets in Chapter III.)

2.6.4. Utilizing Minimum Value "All" Training Dataset

The minimum value "all" training dataset may be necessary for predictors that are more skillful at their minimum value compared to their maximum value. To show that these predictors should be identified in the dataset for use while the other predictors should not be used, the minimum value dataset (without severe weather composite parameters) was plugged into the SWSP. Based on the dendrogram, 30 clusters were chosen for the first cut of the dendrogram. After removing predictors with correlations above 0.4, the top (12) predictors (in order of smallest to largest deviance with Y) were: PMSL2, TEDF1085, DIVG875, HGTTRP, DZDX500, MUPL, U10M, V5SHR, FGEN1000, VVEL30T0, MUCIN, FGEN725 (see Table 2.1).

After deleting predictors with p-values > 0.05, the final model is shown in the R output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.938e+01  1.062e+01   7.474 7.80e-14 ***
TEDF1085     8.040e+01  1.033e+01   7.783 7.10e-15 ***
HGTTRP      -1.334e-04  3.467e-05  -3.849 0.000119 ***
VVEL30T0    -7.995e-01  3.539e-01  -2.259 0.023866 *
DIVG875     -8.684e+03  1.302e+03  -6.670 2.55e-11 ***
PMSL2       -7.768e-04  1.058e-04  -7.344 2.08e-13 ***
MUPL        -1.215e-03  3.817e-04  -3.182 0.001463 **
DZDX500     -2.327e+08  8.957e+07  -2.598 0.009390 **
FGEN1000     1.581e-01  5.961e-02   2.653 0.007973 **
V5SHR        1.585e-02  7.490e-03   2.116 0.034344 *
U10M         9.313e-02  1.755e-02   5.308 1.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2690.6  on 2062  degrees of freedom
Residual deviance: 2130.2  on 2052  degrees of freedom
AIC: 2152.2
```

When validating against the entire dataset (of 2063 cases), the AUROC was 0.7980 and the misclassification error was 23.4%; there were 307 false alarms (431 true nulls) and 176 missed events (1149 true events).

However, using the only predictors’ minimum values within the sample area (162x162 km^2) gave curious results and could have detected areas within the sample area that are not directly related to the severe storm environment. This could also be true for the maximum value dataset because of the large sample area (162x162 km^2), which is another reason why training datasets using a smaller sample area should be investigated (see section 2.7). The only missing predictor in Table 2.3 from the top skillful predictors in Sherburn et al. (2016) was 0-3 km temperature lapse rate (LR03), which is shown near the top of the Table 2.7. The minimum value of LR03 is more skillful than its maximum value. The mean sea level pressure (PMSL and PMSL2) also showed skill in the Sherburn (2018) study, but was not utilized for physical reasoning. It was found that the maximum value of a predictor was more skillful than the

minimum value of the same predictor for only about half of the predictors in the predictor list. Therefore, predictors that maximize skill at their minimum value were identified to be combined with the predictors that maximize skill at their maximum value for a combined "all" training dataset.

It is important to note that the maximum and minimum predictors should be labeled, because they are representing two different extremes within the sample area (e.g., 5x5 grid boxes). When using the maximum value of a predictor, that is stating that the largest value of that predictor in the sample area is being compared between the event and null environments. For example, if a predictor has values ranging from 0 to 10 in the null environments (Y=0), and 2 to 11 in the event environments (Y=1), the maximum value of that predictor would not be as skillful as the minimum value of that predictor. There is a larger difference (discrimination) between the minimum values (0 compared to 2) compared to the maximum values (10 compared to 11), so the "minimum predictor" would be labeled as more skillful. This should not be confused with predictor values increasing or decreasing; for example, increasing CAPE values are associated with increasing chances of severe weather. This example is relating to the CAPE value at a point location increasing or decreasing, not a change in the distribution of CAPE values over an area. This can only be avoided if the average predictor was used (which is not recommended) or the predictor values were taken at a point location (object-based) instead of over an area; see section 6.4. Also, the signs of the coefficients in the probabilistic model may seem counterintuitive because the predictor is increasing/decreasing at the same time the other predictors in the model are increasing/decreasing. A probabilistic model should only be interpreted holistically as one system; see Table 2.13

To help identify predictors that are more skillful at their minimum value compared to their maximum value, the smallest deviance with Y method was used. The deviance with Y was calculated for each predictor in both the maximum value and minimum value datasets, and was compared (Table 2.8). For example, if a predictor had a smaller deviance with Y in the maximum value dataset, the predictor was more skillful at its maximum value compared to its minimum value. If the maximum value of the predictor is more skillful, this does not mean that the predictor is most skillful at its highest value (although this is true in many cases); that will be shown by the sign of the coefficient in the probabilistic model. For example, the maximum value of the SBLI800 predictor within the sample area is more skillful than the minimum value in the sample area, but the coefficient of the SBLI800 predictor in the probabilistic model is negative, which indicates that a decreasing value increases the probability of an event (if all other predictors in the probabilistic model are held constant). The results shown in Table 2.8 led to a new predictor list including each predictor at either its maximum value or minimum value in the sample area (162x162 km^2). Many of the associated relationships are intuitive; for example, a more negative divergence (i.e., convergence) near the surface leads to stronger upward vertical motion, which can indicate stronger storms. The top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR maximum and minimum value "all" dataset (2063 cases) is shown in Table 2.9.

Using this combined dataset as the training dataset for the SWSP gave the following results (see Table 2.17):

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.601e-01  5.343e-01  -1.610 0.107475
min_HGTTRP   -1.493e-04  3.545e-05  -4.212 2.53e-05 ***
min_TEDF1085  5.993e+01  1.029e+01   5.827 5.65e-09 ***
TKE925        2.976e-01  4.093e-02   7.272 3.53e-13 ***
min_FGEN1000  1.470e-01  6.098e-02   2.411 0.015923 *
FGEN725       1.814e-01  4.471e-02   4.057 4.97e-05 ***
min_DIVG875  -5.115e+03  1.353e+03  -3.780 0.000157 ***
UESHR         7.692e-02  8.045e-03   9.560  < 2e-16 ***
NVWSTRP       6.032e+01  1.558e+01   3.872 0.000108 ***
U10M          6.869e-02  1.783e-02   3.852 0.000117 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2690.6  on 2062  degrees of freedom
Residual deviance: 2024.8  on 2053  degrees of freedom
AIC: 2044.8
```

When validating against the entire dataset (of 2063 cases), the AUROC was 0.8231 and the misclassification error was 22.8%; there were 269 false alarms (469 true nulls) and 201 missed events (1124 true events). This was a false alarm rate of 36.4% (at the optimal cutoff) while also having the lowest misclassification error observed out of all the probabilistic models produced so far (see sections 2.7 and 2.8 for discussion). The optimal cutoff of the model was 0.5373.

Adding up the absolute value of the coefficients of the predictors (β's) reveal how much of the outcome (Y) is explained by the predictors in the model; this does not include the intercept (β_0). The β's of the predictors added up to 0.489; the predictors in the model combined contribute 48.9% of the information needed to fully predict Y (i.e., explained variation in Y). The explained variation in Y given by each predictor in the model was the following, in order from largest to smallest contribution to an increase in the probability of an event: 0.09908 (UESHR), 0.07478 (TKE925), 0.07014 (min_TEDF1085), 0.05038 (U10M), -0.04875

(min_HGTTRP), 0.04386 (FGEN725), -0.04264 (min_DIVG875), 0.03470 (NVWSTRP), 0.02441 (min_FGEN1000). UESHR, TKE925, and minimum TEDF1085 gave the largest contributions to the model. This indicated that the u-component (west-east) of the effective bulk wind shear, turbulent kinetic energy at 925 hPa (within the boundary layer), and the minimum 1000 hPa to 850 hPa (surface to around 1.5 km) moist potential temperature lapse rate all contributed to the probability of an event in some way. The effective bulk wind shear is similar to the 0-6 km wind shear, but it accounts for storm depth, so it starts at the effective inflow base (instead of the surface).

It is important to remember that the probabilistic model was created to tell the state of the convective environment while minimizing the repeat of predictive information. In other words, the SWSP was designed to find the combination of skillful predictors that brings skill to the model while also contributes the most original predictive information out of the predictor list. Therefore, the choice of the level of certain predictors may seem odd for some predictors, but it was the level that did not highly correlate with other chosen predictors and it still skillfully discriminated between severe and non-severe (unverified warnings) environments; this draws more predictive information out of the dataset, which can increase the total explained variation in Y.

Calculating the explained variation in Y by the model can be done in an alternative way by using the McFadden's Pseudo-R^2. This gave 0.247, which is equivalent to roughly 50% explained variation in Y by the model (see Step 10), which corroborates with the other calculation. This was a substantial increase in the explained variation of Y compared to the probabilistic model produced using just the maximum "all" dataset (which explained 35.8% of the variation in Y). However, some of the predictors in the combined "all" model (e.g.,

tropopause height, and tropopause vertical wind shear) may be difficult to predict with operational NWP models, so this would need to be tested for real-time forecasting operations. The combined dataset allowed for more skill in the model by maximizing the skill of each predictor in the training dataset, since some predictors are more skillful at their minimum value (e.g., lapse rates which decrease in value with a greater slope in a skew-T diagram). See Table 2.10 for additional model evaluation of this probabilistic model, which provides proof that the SWSP does produce probabilistic models with interpretable probabilities.

2.6.5. Utilizing 3-Hour Difference in Maximum/Minimum Value "All" Training Datasets

The 3-hour difference predictors that were the most skillful based on the smallest deviance with Y were examined; see Section 2.4 for more details on these predictors. The most skillful predictors based on smallest deviance with Y (in order of skill) were: Diffmax_TKE1000, Diffmax_MAXTEVV, Diffmax_W10M, Diffmax_TKE925, Diffmin_DIVG1000, Diffmax_FGEN725, Diffmax_TKE950, Diffmax_FGEN700, Diffmax_DIVG300, Diffmax_TKE900, and Diffmax_V10M. For example, the 3-hour change in the minimum value of divergence at the surface (Diffmin_DIVG1000) within the sample area (162x162 km^2) was a skillful predictor in discriminating between a non-severe (unverified warnings) and severe HSLC convective environment.

Although there were numerous 3-hour difference predictors that came up as significant (p-value less than 0.05) during the model evaluation, the models containing only 3-hour difference predictors were not considered skillful. The significant predictors in the model included: Diffmax_TKE1000, Diffmax_MAXTEVV, Diffmin_DIVG1000, Diffmax_FGEN725, Diffmax_DIVG300, Diffmax_FGEN550, Diffmax_S25MG, Diffmax_TEDF1085,

Diffmax_V6SHR, Diffmin_U6SHR, Diffmax_VVEL30T0, Diffmax_DIVG750. For reference, Diffmax_TKE1000 explained 11.7% of the variation in Y on its own. This value was 11.3% for Diffmax_MAXTEVV, 11.3% for Diffmin_DIVG1000, . . . , 6.8% for Diffmax_DIVG750. And as noted, there were other significant predictors not mentioned. This type of analysis could prove useful for future work. When the significant predictors mentioned above were combined in a model, each predictor explained between 2% to 4.5% of the variation in Y after all other predictors were added to the model, except for DIVG1000 which explained 8.2%.

The maximum value and minimum value "all" datasets were more skillful than the 3-hour difference in maximum/minimum value "all" datasets. However, the sample area of the predictor values (i.e., 162x162 km^2) was considerably large, so these 3-hour difference predictors could be examined within smaller sample areas in future studies. Running the combination dataset containing all datasets (i.e., maximum, minimum, 3-hour difference maximum, 3-hour difference minimum) resulted in the SWSP not selecting any predictors from the 3-hour difference datasets. Most of the chosen predictors in the combination datasets were from the maximum value dataset, and these maximum predictors tended to be the most skillful. Therefore, the results from the 3-hour difference datasets and the combination datasets are not shown.

2.6.6. Utilizing "Tornado" Training Dataset

Note that the "tornado" dataset is a subset of the "all" dataset that includes only tornado cases (event cases) and unverified tornado warnings (NWS false alarms; null cases). The number of observations (cases) is smaller in the "tornado" dataset (865 cases versus 2063 cases), so this may weaken the relationship between the predictors and Y; however, it is still a substantial sample size. (This also means that the deviance and AIC of these "tornado" models cannot be

compared to the “all” dataset models.) The correlation matrix and dendrogram for the maximum value “tornado” dataset is shown in Figure 2.10 for comparisons to the maximum value “all” dataset shown in Figure 2.1. Note how there are once again high correlations outside of the clusters, which is the reasoning for Step 7 in the SWSP.

The top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR maximum and minimum value “tornado” dataset (865 cases) is shown in Table 2.11. Comparing the top most skillful predictors (determined by smallest deviance with Y) between the combined maximum and minimum value “all” (Table 2.9) and the combined maximum and minimum value “tornado” (Table 2.11), predictors depicting moisture (e.g., specific humidity) and low-level shear were more important in discriminating between non-tornadic and tornadic convective environments.

All severe weather composite parameters were kept in the dataset to determine if similar results from section 2.6.3 were found with the “tornado” dataset. Using this combined “tornado” training dataset gave 17 skillful predictors when allowing min_TEDF1050 and U1SHR to stay in the list even though they had correlations around 0.405 with min_SHERB35. This was to see if these two predictors were more useful in the model than min_SHERB35, because they would have been deleted from the list since min_SHERB35 was in their cluster and had the smallest deviance with Y. This was combining the knowledge from before that severe weather composite parameters should not be in the predictor list, which was confirmed again here.

After removing predictors with p-values greater than 0.05 (if the removal improved the AIC of the model), the final predictors and coefficients are shown in the R output:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.949e+00  2.796e+00   2.128  0.03335 *
min_SHERB35   4.269e-01  6.527e-01   0.654  0.51307
min_SPFH1815 -3.534e+02  8.242e+01  -4.287 1.81e-05 ***
U1SHR         8.664e-02  2.423e-02   3.575  0.00035 ***
min_DIVG650  -1.899e+04  4.165e+03  -4.560 5.12e-06 ***
min_TEDF1050  2.828e+02  8.753e+01   3.231  0.00123 **
min_DIVG875  -1.167e+04  2.881e+03  -4.050 5.13e-05 ***
EFFSRH        2.266e-03  9.641e-04   2.351  0.01873 *
NSBCIN       -4.756e-02  2.366e-02  -2.010  0.04438 *
V05SHR        8.354e-02  3.197e-02   2.613  0.00898 **
MUPL         -1.215e-04  3.783e-05  -3.211  0.00132 **
RH2M         -6.870e-02  2.942e-02  -2.336  0.01951 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 782.15  on 864  degrees of freedom
Residual deviance: 526.28  on 853  degrees of freedom
AIC: 550.28
```

The minimum value of SHERB35 (i.e., indicates potential for a HSLC event) was not significant, because most of the cases had a minimum value of zero. SHERB35 most likely had the smallest deviance with Y in its cluster due to the large amount of zero values for this predictor in the dataset (more zero values for null cases than event cases would be considered a distinction between nulls and events). This shows another reason why a severe weather composite parameter may show skill individually but not in a probabilistic model. As mentioned in section 2.6.3, severe weather composite parameters may have missing values or be set to zero for certain situations which diminishes the predictors skill in a probability prediction.

After deleting the SHERB35, the final predictors and coefficients are shown in the R output (see Table 2.18):

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.745e+00  2.519e+00   2.678 0.007415 **
min_SPFH1815 -3.724e+02  7.766e+01  -4.795 1.63e-06 ***
U1SHR         9.213e-02  2.278e-02   4.045 5.23e-05 ***
min_DIVG650  -1.927e+04  4.153e+03  -4.641 3.47e-06 ***
min_TEDF1050  2.984e+02  8.435e+01   3.538 0.000404 ***
min_DIVG875  -1.167e+04  2.887e+03  -4.042 5.31e-05 ***
EFFSRH        2.461e-03  9.156e-04   2.687 0.007202 **
NSBCIN       -4.658e-02  2.379e-02  -1.958 0.050232 .
V05SHR        8.683e-02  3.153e-02   2.754 0.005889 **
MUPL         -1.204e-04  3.769e-05  -3.193 0.001407 **
RH2M         -7.427e-02  2.815e-02  -2.639 0.008326 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 782.15  on 864  degrees of freedom
Residual deviance: 526.71  on 854  degrees of freedom
AIC: 548.71
```

Removing NSBCIN from the model increased the AIC to 551.32 so it was kept in the model. When validating against the entire dataset (of 865 cases), the AUROC was 0.8727 and the misclassification error was 12.4%; there were 83 false alarms (62 true nulls) and 24 missed events (696 true events). This was a false alarm rate of 57.2% at the optimal cutoff (see sections 2.7 and 2.8 for discussion). The optimal cutoff of the model was 0.4798.

The explained variation in Y given by each predictor in the model was the following, in order from largest to smallest contribution to an increase in the probability of an event: 0.07252 (U1SHR), -0.06723 (min_SPFH1815), 0.06194 (min_TEDF1050), -0.05900 (min_DIVG650), -0.05429 (min_DIVG875), 0.03969 (V05SHR), -0.03359 (MUPL), -0.03243 (RH2M), 0.02335 (EFFSRH), -0.02010 (NSBCIN). The predictors in the model combined contributed 46.4% of the information needed to fully predict Y (i.e., explained variation in Y). The three predictors that contributed the most to the model were the u-component (west-east) of the surface to 1 km wind

shear (U1SHR), minimum value of the specific humidity (moisture) around 150 hPa to 180 hPa above ground level (SPFH1815), and the minimum value of the moist potential temperature lapse rate from the surface to mid-levels (TEDF1050). For reference, specific humidity is essentially the amount of moisture in a system based on mass (of the moisture compared to the mass of the system), according to the American Meteorological Society definition.

Because the probabilistic model works as a holistic system, the signs of the coefficients cannot be interpreted individually. In general, the chosen predictors can be related to tornadic mesoscale environments. For example, increasing low-level wind shear (U1SHR and V05SHR) increases the chance of a tornadic environment. An increasing eastward wind shear (U1SHR) right above an increasing northward wind shear (V05SHR) in the 0-1 km could represent a clockwise turning of winds with height (veering wind) in the lower levels. This lower-level directional shear increases near-surface rotation, which could help the development of a rotating updraft. This is important at lower levels (0-1 km) in typical HSLC storms in the SEUS, which are shallower and closer to the ground, compared to HSHC storms in the U.S. Plains. The moist potential temperature lapse rate from 1000 hPa to 500 hPa (TEDF1050) indicates the stability of the environment with decreasing values associated with severe environments. If the moist potential temperature is decreasing with height (negative lapse rate), the atmosphere is unstable (with lift and saturation present), which is the type of environment that produces severe storms.

It should also be kept in mind that the NARR horizontal grid spacing is 32 km and the maximum or minimum value of each predictor is chosen within a 162x162 km^2 area centered over the case location, so the predictor values are representing a mesoscale environment over a large area. The skillful predictors in the “tornado” dataset could be representing a more severe convective environment compared to the “all” dataset, so it is not necessarily representing the

variables needed for tornadogenesis which is at the storm-scale. This is one of the reasons for developing a new training dataset that has higher resolution data (see Chapter III). Also, there were substantially more event cases than null cases in the "tornado" dataset, so more null cases may also be needed for further research.

2.6.7. Utilizing "Severe Wind Speed" Training Dataset

Note that the "severe wind speed" dataset is a subset of the "all" dataset that includes only severe wind speed cases (event cases) and unverified severe thunderstorm warnings (NWS false alarms; null cases). See section 2.3 for more information. The number of observations (cases) is smaller in the "severe wind speed" dataset (1198 cases versus 2063 cases), but still a substantial sample size. (This also means that the deviance and AIC of these "severe wind speed" models cannot be compared to the "all" dataset models.) The top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR maximum and minimum value "severe wind speed" dataset (1198 cases) is shown in Table 2.12.

Comparing the top most skillful predictors (determined by smallest deviance with Y) between the combined maximum and minimum value "all" (Table 2.9) and the combined maximum and minimum value "severe wind speed" (Table 2.12), predictors depicting low-level turbulent kinetic energy (e.g., TKE950) and the product of moist potential temperature lapse rate and pressure vertical velocity (omega) in the lower to mid-levels (e.g., TEVV2K) were more important in discriminating between non-severe and severe convective environments for severe wind speed cases. In general, low-level TKE is associated with large low-level temperature lapse rates and low-level wind shear, which are associated with severe convective environments.

However, the low-level TKE predictors were parameterized in the Eta model (NARR) and were dependent on model configuration, so these values may not well represent the observed conditions. Model configuration in general is not a concern, as long as the probabilistic models are trained and operated using the same gridded (NWP) model, which is recommended. As noted previously, since the TEVV predictors combine more than one predictor, these are considered composite parameters, so it may be a good idea to take them out of the predictor list in future work for the multiple reasons described throughout section 2.6. However, the TEVV predictors were calculated by Sherburn et al. (2016) to approximate the release of potential instability, which showed high skill in their study, so they were kept in the predictor list for Chapter II. Of course, surface level wind (W10M, V10M, and U10M) is near the top of the skillful predictor list, because the event cases are reports of severe wind speeds (i.e., wind gust of 65 knots and faster), but this is a good rationality check on the quality of the dataset and predictor selection methods.

All severe weather composite parameters were kept in the combined dataset to determine if similar results from section 2.6.3 and 2.6.6 were found with the "severe wind speed" dataset. The correlation matrix and dendrogram for the 13 chosen predictors is shown in Figure 2.11 to provide an example a final correlation matrix given by the SWSP. The list of chosen predictors in order of skill (smallest deviance with Y) were: TKE925, min_DIVG875, FGEN700, min_TEDF1085, U10M, min_DZDX500, min_MUPL, min_HGTTRP, min_FGEN1000, min_MUCIN, min_U6SHR, min_LR75 (should be labeled maximum LR75 due to the sign error in the dataset), VVEL975.

The logistic regression step on this chosen predictor list resulting in the results shown:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.901e+00  1.356e+00  -1.402 0.160982
min_HGTTRP   -5.796e-05  4.965e-05  -1.167 0.243096
min_TEDF1085  5.373e+01  1.427e+01   3.766 0.000166 ***
VVEL975      -5.257e-01  5.376e-01  -0.978 0.328153
min_FGEN1000  2.071e-01  8.287e-02   2.499 0.012468 *
TKE925        4.043e-01  5.065e-02   7.984 1.42e-15 ***
FGEN700       2.187e-01  5.514e-02   3.965 7.33e-05 ***
min_DIVG875  -5.474e+03  1.733e+03  -3.158 0.001587 **
min_DZDX500  -4.766e+08  1.171e+08  -4.070 4.70e-05 ***
min_MUCIN     4.897e-05  1.818e-03   0.027 0.978513
min_MUPL     -1.492e-03  5.663e-04  -2.636 0.008398 **
min_LR75      1.707e-01  1.581e-01   1.080 0.280304
min_U6SHR    -3.157e-02  9.851e-03  -3.204 0.001353 **
U10M          5.695e-02  2.268e-02   2.511 0.012046 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1660.7  on 1197  degrees of freedom
Residual deviance: 1224.2  on 1184  degrees of freedom
AIC: 1252.2
```

Deleting predictors with p-values greater than 0.05 lowered the AIC except for min_HGTTRP so it was kept in the model. After deleting the insignificant predictors, the final predictors and coefficients are shown in the R output (see Table 2.19):

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.037e-01  6.576e-01  -0.918 0.358594
min_TEDF1085  5.532e+01  1.411e+01   3.920 8.84e-05 ***
min_HGTTRP   -7.933e-05  4.504e-05  -1.761 0.078189 .
min_FGEN1000  2.144e-01  8.337e-02   2.572 0.010107 *
TKE925        3.935e-01  5.006e-02   7.860 3.85e-15 ***
FGEN700       2.106e-01  5.389e-02   3.908 9.31e-05 ***
min_DIVG875  -4.782e+03  1.659e+03  -2.882 0.003950 **
min_DZDX500  -4.660e+08  1.166e+08  -3.998 6.39e-05 ***
min_MUPL     -1.441e-03  5.601e-04  -2.572 0.010097 *
min_U6SHR    -3.275e-02  9.779e-03  -3.349 0.000812 ***
U10M          5.314e-02  2.213e-02   2.401 0.016357 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1660.7  on 1197  degrees of freedom
Residual deviance: 1226.7  on 1187  degrees of freedom
AIC: 1248.7
```

When validating against the entire dataset (of 1198 cases), the AUROC was 0.8287 and the misclassification error was 23.8%; there were 157 false alarms (436 true nulls) and 128 missed events (477 true events). The optimal cutoff of the model was 0.4696. This was a false alarm rate of 26.5% at the optimal cutoff (see sections 2.7 and 2.8 for discussion). However, severe thunderstorm warnings are issued for other types of severe weather hazards as well, so if the null cases were trimmed to only severe thunderstorm warnings issued for a severe wind speed hazard, this probabilistic model could improve. The case list definitions are important for model skill.

The explained variation in Y given by each predictor in the model was the following, in order from largest to smallest contribution to an increase in the probability of an event: 0.12681 (TKE925), 0.06991 (min_TEDF1085), 0.06203 (FGEN700), -0.05272 (min_DZDX500), -0.04663 (min_DIVG875), -0.04417 (min_U6SHR), 0.03908 (U10M), 0.03029 (min_FGEN1000), -0.02978 (min_HGTTRP), -0.01760 (min_MUPL). The predictors in the model combined contribute 51.9% of the information needed to fully predict Y (i.e., explained variation in Y). The near-surface turbulent kinetic energy (TKE925) is the largest contributor to the model, which as previously stated, could be swapped with a related predictor (other than TKE) since TKE is not accurately represented in the NARR.

The minimum u-component of the 0-6 km wind shear (U6SHR) was also in the model. When studying the differences between rapidly and slowly developing mesoscale convective systems (MCSs), Coniglio et al. (2010) found a smaller 3 km to 10 km vertical wind shear in the environment of rapidly developing MCSs compared to slowly developing. This brings up a good point that the stage of storm development will dictate which environmental predictors and at what strengths are most important during that time. This will have to be kept in mind during

future probabilistic modeling studies. In general, a decreasing moist potential temperature lapse rate in the lower levels (TEDF1085) is associated with severe wind speeds, which are mostly straight-line wind events in the HSLC cases. Fronts play a role in straight-line wind events and this is evident in the contributions of the surface frontogenesis (FGEN1000) and around 3 km frontogenesis (FGEN700) predictors in the model. Because severe wind speeds occur at the storm-scale, to predict the probability of a severe wind speed using gridded (NWP) model data with a horizontal grid spacing around 32 km (NARR), other predictors (besides surface wind speeds) are necessary in the probabilistic models. Mesoscale probabilistic models can be used in conjunction with other severe weather forecasting tools to provide additional assistance when making decisions on severe thunderstorm warnings, which include severe wind speeds.

U10M is the u-component (west-east) of the surface wind speed, which was chosen over the v-component (north-south) of the surface wind speed (V10M) and the total surface wind speed (W10M). This is because U10M brought more original predictive information to the model and/or had the lowest deviance with Y. During examination it was found that U10M had a higher deviance than W10M or V10M, but brought more original predictive information to the model (lower correlation among predictors). W10M and V10M had a correlation of 0.845 (same cluster) and they were highly correlated with other predictors (same cluster with TKE925, which had a lower deviance). For reference, V10M and TKE925 had a correlation of 0.702, and W10M and TKE925 had a correlation of 0.763. This is an example of how a skillful predictor (e.g., W10M) may not be chosen in the probabilistic model. The scale of the convective environment being represented in the training dataset is important, because the probabilistic models trained with the NARR data (mesoscale data) should not be used for storm-scale prediction. This can be

seen with the chosen predictors in the "severe wind speed" probabilistic model, which may not make sense at the storm-scale.

The HSLC severe weather composite parameters (i.e., all versions of MOSH and SHERB) were not a chosen predictor in the models using either the "tornado" or "severe wind speed" datasets, which gives evidence that more specific training datasets (i.e., separating datasets by severe weather hazard type) may provide improvements in severe wind speed prediction and tornadogenesis prediction, separately. The HSLC severe weather composite parameters were developed using the "all" dataset (or similar), so the focus of those composite parameters was predicting HSLC severe convective environments in general. It was determined that a higher resolution data and/or more specific training datasets may be needed to produce more skillful results in the prediction of a specific severe weather hazard type. Separating the cases into different datasets lowers the number of cases (sample size) in the training dataset, which can lower the skill of the model, so this should be done with caution.

2.7. Limitations to Results

It is possible some variables (predictors) that represent atmospheric processes associated with tornadogenesis and severe wind speed development are either missing or not accurately represented in the NARR gridded output. This can be true for any gridded model output, so it should always be considered a limitation. During statistical analysis of the datasets, it was roughly estimated that up to half of the information necessary to fully represent these processes was available in the most skillful probabilistic model produced using the entire NARR training dataset, which included the subsequent predictors calculated with the raw NARR output. This was determined by examining the percentage of explained variation in Y (where Y=1 is an event

case) provided by the predictors in each skillful probabilistic model (i.e., the predictors were centered and scaled to reveal the contribution of the predictors in the model). In the maximum/minimum combined "all" probabilistic model (section 2.6.4), the predictors combined contributed up to 48.9% of the information needed to fully predict Y (i.e., explained variation in Y).

The case list used in this research was also used to develop the MOSH and MOSHE (Sherburn et al. 2016), and it is important to be aware of the limitations of this dataset. The definition of the null cases in this dataset are essentially based on the subjectivity of the forecaster issuing the severe warning and the lack of observed storm reports. The explicit definition of the null cases, which are defined in Sherburn and Parker (2014), are: "the initial latitude–longitude point [as determined by the NWS operational meteorologist] of a severe thunderstorm or tornado warning that was issued in an HSLC environment [defined in section 2.3] when there were no severe reports from the Storm Data archives in the corresponding CWA [NWS county warning area] throughout that convective day (1200–1200 UTC)." The event cases also have a specific definition: event cases in the dataset are NWS official storm reports were EF1 or greater tornado reports and severe wind speed reports (wind gust $\geq$ 65 knots). For further discussion on the case list definitions and reasoning, see section 2.3. Because a severe thunderstorm and/or tornado warning was issued for all null cases, it is safe to assume that all cases occurred in convective environments, which is why it is stated that the resulting probabilistic models discriminate between HSLC "convective environments." These specific event and null definitions allow us to focus on the bigger impact events (stronger severe wind speeds and tornadoes) in HSLC convective environments which were difficult to forecast by experienced NWS operational forecasters. Therefore, the models that are produced using this

case list discriminate between two environments which appeared to be severe convective environments but may have had slight differences that are difficult to forecast and/or unknown (statistically) significant differences.

Since the case definitions used in this study depend on the storm reports existing in the NOAA Storm Data archive (NOAA NCEI 2019), it is important to know the limitations of this data archive. This includes discrepancies in the reports due to poor characterization of the scope and magnitude of estimated severe wind speeds (e.g., Doswell et al. 2005, Trapp et al. 2006, Smith et al. 2013), complexities of the damage-to-wind speed relationship (e.g., Doswell et al. 2009), subjectivity in tornado report severity and count (e.g., Verbout et al. 2006), primary convective mode that produces severe wind speeds in region (e.g., Smith et al. 2013), population bias and time of day (e.g., King 1997, Brooks et al. 2003, Trapp et al. 2006), and spatial distance between observed reports and actual reports (e.g., Sobash et al. 2011). The incompleteness and uncertainties of the case list, due to using the Storm Data archive, limit what the models predict. With the case list used for the probabilistic models in this research, these models predict a probability of a severe weather hazard (wind gust $\geq$ 65 knots and/or EF1+ tornado) in HSLC convective environments where a severe warning was issued (which is subjective). This is based on a partial representation of severe storm reports that occurred in HSLC environments in the SEUS (i.e., the severe weather hazards that were reported), which affects the event and null case list. The spatial and temporal definitions depend on the training dataset; see section 4.4. For further discussion on the case list definitions and reasoning, see section 2.3.

Therefore, the MOSH, MOSHE, and probabilistic models created from this dataset need to be used carefully and with all data definitions and limitations in mind. Also, this research focused on reducing FAR, because it did not address the increase in POD which needed a larger

amount of missed cases in the dataset. To develop a severe weather probabilistic model that is not just based on NWS-warned cases requires developing an additional database that includes a substantial number of NWS missed events; that database would be able to address the POD forecasting issue. The specific interpretation of the probabilistic models is based on the training dataset used to produce the model. This includes the case definitions (e.g., event and null definitions), gridded model used to choose the predictors and coefficients (e.g., RAP), predictor values (e.g., maximum value), sample size (e.g., sample radius), sample time (e.g., difference between observation and model time), and data type (e.g., analysis, forecast). These specifications will also determine the spatial and temporal specifications of the prediction. That is, the probability of the severe weather hazard is valid within # km and # hours; see section 4.4 for more details. Interpretation of the probabilistic models is discussed further throughout the dissertation, including Chapter I, and sections 2.3, 4.4, 5.3, 5.4, and 6.2.

There are also possible spatial and temporal uncertainties in the official storm reports (i.e., reported location and time of event cases), specifically, discrepancies between when (time) and where (location) the severe weather hazard occurred versus the time and location recorded in the official storm report. The reported severity of the severe weather hazard (e.g., EF1 versus EF2) also includes uncertainty. These uncertainties exist because the individual NWS Weather Forecasting Office (WFO) that forecasts for that region of interest oversees the estimation of the location, time, and severity of the reported severe weather hazard, which introduces subjective human error into the reports. Yet, these uncertainties are relatively negligible in this research study, and the official storm reports are the best observations of tornadoes and severe wind speeds available. The spatial uncertainties in the event cases are not a concern since areas encompassing the storm report (i.e., 5x5 grid boxes) are sampled, not point locations. There are

similar issues when estimating the location of the null cases; however, accuracy of the point location of the null case is not necessary since we are concerned with sampling an area of any HSLC null environment. The null cases did not have an observed storm report of any kind within the county warning area or adjacent county warning areas (filling the sample area of the null environment). The temporal uncertainties for event and null cases are minor since the possible discrepancy in timing most likely would not affect which gridded (NWP) model output time is sampled, so the sample most likely would not change based on these discrepancies. The severity uncertainty is not a concern because cases are not evaluated by severity, that is, all tornado cases are in the "tornado" dataset, and all severe wind speed cases are in the "severe wind" dataset. Lastly, EF0 tornado reports, which could possibly be misidentified as severe wind speed reports, are not included in either dataset.

Through the process of model validation (i.e., section 2.6, Step 10), it was discovered that the master case list had spatial and temporal correlation errors among the events and the nulls due to various issues. This can cause p-values to be lower or higher than the actual value. For example, there were multiple event cases that reported a tornado at the same model time and within the same sample area, which may have been the same tornado in slightly different locations as it moved across the area. This also occurred with severe wind speed event cases. There were null cases that occurred at the same model time and within the same sample area as well, which was due to adjacent CWAs reporting severe warnings at the same time. No matter the reason, if cases occurred at the same model time and within the same sample area, they will have the same values for all predictors; this is a duplicate sample in the dataset, which can cause correlation errors. Therefore, the dates and times of the cases can be manually compared, and duplicate samples (i.e., a case where all predictor values are the same as another case) will be

deleted from the training datasets in further chapters. Another concern was that the severe weather hazards (event cases) in the master case list occurred up to 1.5 hours before or after the NARR model time (i.e., sample time of predictor values), which is not ideal when studying severe convective environments that can change rapidly. This time discrepancy also led to the inclusion of sample areas that did not accurately sample the local storm environment because the storms were not located in the sample area (i.e., 162x162 km^2) at the sample time (i.e., NARR output time). The NARR data is output at 29 vertical pressure levels, so it has a larger vertical spacing (i.e., 50 hPa instead of 25 hPa) in the output between 700 hPa and 300 hPa. Although unlikely, this may limit the potential skill of predictors within those missing levels.

The results in Chapter II are important because they set up the basic procedure to produce workable probabilistic models, but more steps are needed with the training dataset to improve these models. To improve these models, new data collection and processing methods need to be developed to minimize the errors and issues outlined above. The errors and issues found in Chapter II motivated the work completed in Chapters III and IV. Because the predictors that represent atmospheric processes associated with tornadogenesis and severe wind speed development may not be accurately represented in the training dataset due to the large sample area (162x162 km^2) surrounding the case location used to calculate the (maximum or minimum) predictor values, and/or due to the large horizontal grid (32 km) and temporal (3-hourly) spacing of the NARR model output, this led to two hypotheses: 1) the probabilistic models will be improved by using smaller sample areas to calculate the maximum and minimum of predictor values in the training dataset, and 2) the probabilistic models will be improved by using higher resolution gridded model data (i.e., gridded model with smaller grid spacing and temporal spacing) for the predictor values in the training dataset. Predictors that are necessary for severe

weather hazard (specifically tornado and/or severe wind speed) prediction may not be sufficiently and/or currently represented in the variable list (Table 2.1), so this led to two more hypotheses: 3) predictors that are known to be necessary for tornado and/or severe wind speed prediction are not sufficiently represented or existing in the variable list (and need to be added to gridded model output), and 4) there are predictors beyond our current understanding that are necessary for tornado and/or severe wind speed prediction. If Hypotheses 1-3 are proven to be false, this would provide evidence for Hypothesis 4 by process of elimination. Based on numerous weather forecasting research studies, there are variables (predictors) that are not well represented in gridded models due to the parameterization schemes that calculate the predictor values, and there are also potentially skillful predictors that are not included in operational gridded (NWP) models (e.g., quantifying cold pool strength). The "missing" skill in the probabilistic models may be due to one or more of the reasons outlined above. Since Hypothesis 3 (NWP modeling study) and Hypothesis 4 (test relying on results of Hypothesis 3) were beyond the scope of this dissertation study, we focus on Hypotheses 1 and 2 in subsequent chapters.

2.8. Summary

A severe weather statistical procedure (SWSP) was constructed using a combination of statistical techniques to identify statistically significant differences between two types of convective environments (i.e., a severe weather hazard occurring in one environment versus the severe weather hazard not occurring in the other environment), using a binary response variable. The SWSP was designed specifically for large datasets of highly correlated meteorological data and can create a probabilistic model within minutes. High-shear, low-CAPE (HSLC) tornado and severe wind speed (event) cases and unverified NWS warning (null) cases within the

Southeastern United States (SEUS) were used to test the statistical procedure and create probabilistic models, which provided a proof of concept for the SWSP. Predictor values were taken from North American Regional Reanalysis (NARR) gridded model data. The probabilistic models produced by the SWSP provide interpretable probabilities of a severe weather hazard, which can provide an easily understandable message during emergency communications.

Proof that the SWSP produce probabilistic models that provide interpretable probabilities was given in Table 2.10. For example, the probabilistic model predicted a probability between 90% to 100% for 391 cases, and 368 of those cases were an actual event case (Y=1), which equates to 94.12%. If the model predicted perfectly, the average actual probability for the 90% to 100% range is 95%; therefore, the model performed very well for that prediction range. Note that this is not related to the optimal cutoff (probability threshold) for the model; the probability threshold just determines whether to mark a prediction as a null or an event during model evaluation for model comparisons. These results show proof that the skillful probabilistic models give an interpretable probability. An interpretable probability means that if the model predicts there is a 50% probability of an event, an event will occur, on average, 50% of the time. If the dataset (number of cases) is infinitely long, this is theoretically true. This result addresses communication concerns (i.e., false alarms). This model evaluation (Table 2.10) should be completed for each probabilistic model to explain the uncertainty of the predicted probability in more detail.

The same probabilistic model used in Table 2.10 (i.e., the combined NARR maximum and minimum value “all” dataset shown in section 2.6.4) was utilized again in Table 2.13 to give an illustrative example of how a probabilistic model works as a holistic system. Because a predictor in the model is likely to increase/decrease at the same time of an increase/decrease of

another predictor in the model, the odds-ratios of a predictor (calculated when all other predictors in the model are held constant) may not give useful information. It is best to examine the probabilistic model as one complete system. In this example, three cases were randomly chosen (in categories); case 1 was a non-severe (unverified warning) environment (Y=0, predicted probability of 0.104), case 2 was a "more ambiguous" severe environment (Y=1, predicted probability of 0.600), and case 3 was a severe environment (Y=1, predicted probability of 0.983). Comparing between Case 1 and Case 3, the severe case had a lower tropopause height, higher TKE near the surface, higher u-component surface wind speed, greater convergence near the surface, and stronger u-component effective wind shear; these are associated with severe convective environments. Even though TKE is not well represented in the gridded data, the parameterized value is still valuable in a probabilistic model. The difference between frontogenesis at the surface and frontogenesis at 725 hPa (vertical gradient) increased substantially with the severe case, and this was also seen between Case 1 and Case 2. The potential temperature lapse rate increased slightly with the severe case (Case 3 compared to Case 1), which is counterintuitive; however, the other predictors in the model (greater total weight) essentially compensated for this and it should be kept in mind that the training dataset was a parameter subset of lower instability cases. Looking at the Case 2 values, the 0.6 probability prediction, shows how the probability can be dominated by some of the predictors. More predictors in Case 2 are counterintuitive (comparing to Case 1), and other predictors in the model partially compensate for these. Because of the ambiguity, Case 2 was predicted to have a lower probability, near 0.5, which is essentially a tossup.

It was shown that the probabilistic models can predict low false alarm rates (i.e., false positives) within the dataset. However, the severe wind speed (gusts) equal to and greater than

65 knots and tornadoes greater than EF0 are only part of the severe weather hazards that prompt a severe thunderstorm warning, so the other hazards (including hail) need to be included in future studies. Reducing the FAR is an operational forecasting problem for HSLC severe convective environments, so this is a focus in this research. One of the most skillful probabilistic models trained from the “all” dataset predicted a 36.4% false alarm rate (FAR) of a severe weather hazard (at the optimal cutoff) in the dataset. This means when setting the model at the probability threshold that correctly predicted the most cases (optimal cutoff), 63.6% of the null cases were correctly predicted in the dataset. In other words, increasing the true nulls, and thereby decreasing the missed nulls in the dataset, decreases the FAR in the dataset. Therefore, the probabilistic model can discriminate between two HSLC convective environments (when a warning has been issued), where tornadoes and/or severe wind speeds were observed in one environment (events) but not the other environment (nulls). Keep in mind the case definition assumptions stated in section 2.3.

The FAR of a severe wind speed was 26.5% at the optimal cutoff; the definition of a severe wind speed used for the training dataset was a wind gust of 65 knots or greater. Severe thunderstorm warnings are issued if there is a prediction of wind speeds of at least 50 knots, hail at least 1" in diameter, and/or a tornado. Generally, if hail or tornado reports are verified, there was also a verified wind speed greater than 50 knots within the area, so a model that can lower the FAR for severe wind speeds may also lower the FAR for severe thunderstorm warnings. There was a 15-knot increase in the wind speed (gust) threshold for this research; the threshold of 65 knots was used to ensure the wind reports in the training dataset were indeed severe wind speeds, which are difficult to estimate through storm damage reports. These results show a potential for the use of the SWSP in producing probabilistic models that may lower FAR in

NWS operational warnings if the other hazards (in severe weather warnings) are included in the training datasets. In the combined "all" probabilistic model (section 2.6.4), the predictors combined contributed up to 48.9% of the information needed to fully predict Y (i.e., explained variation in Y).

These were promising results for the SWSP and more work is needed to improve the percentage of explained variation in Y by the probabilistic models. When using the NARR (32 km, 3-hourly) training data, the SWSP produced mesoscale probabilistic models that can be used in conjunction with other severe weather forecasting tools to provide additional assistance when making decisions on severe warnings in the near-term. The results produced using the NARR training datasets have theoretical importance and help describe the differences between severe and non-severe (unverified warnings) convective environments. However, the practical use of the chosen predictors in the probabilistic models needs to be investigated further. Looking at the development of the severe weather composite parameters, the predictors were all designed to indicate severity and/or distinction between severe or non-severe environments. They were not designed for predicting the probability of a severe weather hazard (e.g., tornado), which was seen during testing. These severe weather composite parameters were made up of components (predictors) that were already included in the predictor list, so they provide the same predictive information as other predictors in the training dataset. This gives evidence that severe weather composite parameters are not necessary and are not recommended for inclusion in the training datasets that are used for producing probabilistic models, if predictors related to the components of the composite parameter can be included individually in the training dataset.

It was discovered that the training datasets, which consisted of the same cases from Sherburn et al. (2016), produced probabilistic models with spatial and temporal correlation errors

due to various reasons, as discussed in section 2.7. To minimize these correlation errors, new training datasets should be produced using new data collection and processing methods (as described in Chapter III) to create more skillful probabilistic models (e.g., Chapter IV). The new training datasets (Chapter III) should also be designed to be more specific, by focusing in on the local storm environment. The SWSP will not need to change for these improvements to be made. To test if higher resolution gridded (NWP) data produces more skillful models, data from the Rapid Refresh (RAP) model should be used to produce probabilistic models for operational forecasting (as done in Chapter V). The models produced in Chapter II are still valuable as they show skillful predictors that can be useful in predicting tornadoes and/or severe wind speeds within low-instability convective environments. These combinations of statistically significant, weighted predictors are weakly correlated among themselves in each model, which minimizes the overlap of predictive information contributed from each predictor. Some of these predictors have shown better skill in probabilistic prediction, both individually and together, compared to severe weather composite parameters currently used by the NWS (e.g., STP). These results also show the importance of conducting additional research in discriminating between severe and non-severe convective environments in low-instability (HSLC) scenarios.

Lastly, the SWSP has proven to be an easy analysis tool that helps the user quickly analyze a large set of atmospheric variables (predictors) for skill in distinguishing between a non-severe (unverified warnings) and severe convective environment. Multiple types of analysis were described in this chapter, all of which can provide useful results on the predictive information given by each predictor, on its own and in combination with other predictors. If the user seeks interpretable probabilities, the SWSP has proven to produce skillful probabilistic models which can give these interpretable probabilities. With carefully chosen skillful predictors,

the logistic regression will produce a probabilistic model that gives an interpretable probability of an event. The SWSP can be used in future work to find new skillful predictors for use in any severe weather forecasting model.

Table 2.1. List of the pre-selected variables included in the NARR training datasets. The NARR datasets were provided by Sherburn et al. (2016). There are 262 predictors. *- indicates variables were calculated/derived using NARR output. #- indicates some of these variables were calculated/derived using NARR output. - indicates raw NARR output variable. Note that AGL means above ground level.

- SPFH1000, etc.: Specific humidity (kg/kg; 1000 hPa, 975 hPa, ... 500 hPa)

- TMP1000, etc.: Temperature (K; 1000 hPa, 925 hPa, 850 hPa, 700 hPa, 500 hPa)

- VVEL1000, etc.: Pressure vertical velocity (omega) (Pa/s; 1000 hPa, 975 hPa, ... 500 hPa)

- TKE1000, etc.: Turbulent kinetic energy (J/kg; 1000 hPa, ..., 850 hPa)

*- DIVG1000, etc.: Divergence (s^{-1}; 1000 hPa, ..., 100 hPa)

*- FGEN1000, etc.: Frontogenesis (K/[100 km*3h]; 1000 hPa, ..., 500 hPa)

- MDIV850, MDIV30T0, MDIV6030, etc.: Horizontal moisture divergence (kg/kg/s; 850 hPa, 30-0 hPa AGL, 60-30 hPa AGL, 90-60 hPa AGL, 120-90 hPa AGL, 150-120 hPa AGL, 180-150 hPa AGL)

- SPFH30T0, SPFH6030, etc.: Specific humidity (kg/kg; 30-0 hPa AGL, 60-30 hPa AGL, ..., 180-150 hPa AGL)

- VVEL30T0, VVEL6030, etc.: Pressure vertical velocity (omega) (Pa/s; 30-0 hPa AGL, 60-30 hPa AGL, ..., 180-150 hPa AGL)

*- SRH05, SRH1, SRH15, etc.: Storm-relative helicity (m^2/s^2; 0-0.5 km, 0-1 km, ..., 0-3 km)

#- CAPE05, CAPE1, CAPE15, etc., NMLCAPE, NSBCAPE, MLCAPE, MUCAPE: CAPE (J/kg; 0-0.5 km, 0-1 km, ..., 0-3 km, mixed-layer, surface-based, mixed-layer [calculated by KS], most unstable [calculated by KS])

#- MLCIN, MUCIN: CIN (J/kg; mixed-layer, surface-based, mixed-layer [calculated by KS], most unstable [calculated by KS])

*- SBLI500, SBLI550, etc., LIMIN, LIMAX: Lifted index (K; surface-500 hPa, surface-550 hPa, ..., surface-850 hPa, maximum, minimum)

*- SBLFC, MULFC: LFC (m; surface-based, most unstable)

*- TMPLFC: LFC temperature (K)

- NHGTCLB: Cloud base height (m)

*- SBLCL: LCL (m; surface-based)

*- TMPLCL: LCL temperature (K)

*- EFFTOP: Effective inflow layer top (m)

Table 2.1. (continued).

*- EFFBOT: Effective layer bottom (m)

*- EFFSDPTH: Effective storm depth (m)

*- EFFIDPTH: Effective inflow depth (m)

*- EFFNCAPE: Effective normalized CAPE (J/kg/km)

*- EFFSHR: Effective bulk shear magnitude (m/s)

*- UESHR: U-component of effective bulk shear vector (m/s)

*- VESHR: V-component of effective bulk shear vector (m/s)

*- EFFSRH: Effective storm-relative helicity (m^2/s^2)

*- TED1050, TED1070, TEDF02, TEDF025, etc.: Theta-E lapse rate (moist potential temperature lapse rate) (K/m; 1000-500 hPa, 1000-700 hPa, 1000-850 hPa, 850-700 hPa, 0-2 km, 0-2.5 km, ..., 0-6 km)

*- STP: Significant tornado parameter

*- SCP: Supercell composite parameter

*- EHI: Energy helicity index

*- CBSS: Craven-Brooks significant severe parameter

*- VGP: Vorticity generation parameter

*- SHERB3: SHERBS3 (original)

*- SHERBE: SHERBE (original)

*- LR005, LR01, etc.: Temperature lapse rate (K/km; 0-0.5 km, 0-1 km, ..., 0-3 km)

*- LR36, LR35, LR75: Temperature lapse rate (K/km; 3-6 km, 3-5 km, 700-500 hPa)

*- U6SHR, U5SHR, etc.: U-component of shear vector (m/s; 0-6 km, 0-5.5 km, ..., 0-0.5 km)

*- V6SHR, V5SHR, etc.: V-component of shear vector (m/s; 0-6 km, 0-5.5 km, ..., 0-0.5 km)

*- S6MG, S5MG, etc.: Shear vector magnitude (m/s; 0-6 km, 0-5.5 km, ..., 0-0.5 km)

- NVWSTRP: Tropopause vertical wind shear (s^{-1})

- TMP2M: 2-m temperature (K)

- DPT2M: 2-m dew point (K)

- RH2M: 2-m relative humidity (%)

- U10M: U-component of 10-m wind speed (m/s)

- V10M: V-component of 10-m wind speed (m/s)

Table 2.1. (continued).

- THSFC: Surface potential temperature (K)
- PMSL: Mean sea-level pressure (Pa)
- PMSL2: Mesinger mean sea-level pressure (Pa)
- HGT0DEG: Freezing level height (m)
- HPBLSFC: Planetary boundary layer height (m)
- HGTTRP: Tropopause height (m)
- MUPL: Most unstable parcel lifted level (m)
- MUEL: Most unstable parcel equilibrium level (m)
- RHHL1: Hybrid level 1 relative humidity (%)
- PWAT: Precipitable water (kg/m^2)
- CLTPRS, CLBPRS: Cloud-top pressure and cloud-base pressure (Pa)
- PRATE: Precipitation rate [$kg/m^2/s$]

*- VV2, VV25, etc.: Pressure vertical velocity (omega) (Pa/s; 2 km, 2.5 km, ..., 6 km)

*- TEVV5, TEVV7, TEVV8: 1000-500hPa, 1000-700hPa, 1000-850hPa Theta-E lapse rate * vertical velocity product

*- TEVV2K, TEVV25K, etc., MAXTEVV, MINTEVV, TEVVDF: Theta-E lapse rate * vertical velocity product ([Pa*K]/[km*s]; 2 km omega * 0-2 km theta-E lapse rate, 2.5 km ", ..., 6 km ", maximum, minimum, maximum-minimum)

*- W10M: 10-m wind speed (m/s; including both u and v components)

*- MOSH: MOSH (i.e., modified SHERB)

*- MOSHE: MOSHE (i.e., modified SHERBE)

*- SHERB35: SHERBS3 (new, using 3-5 km lapse rate instead of 700-500 hPa)

*- SHERBE35: SHERBE (new, using 3-5 km lapse rate instead of 700-500 hPa)

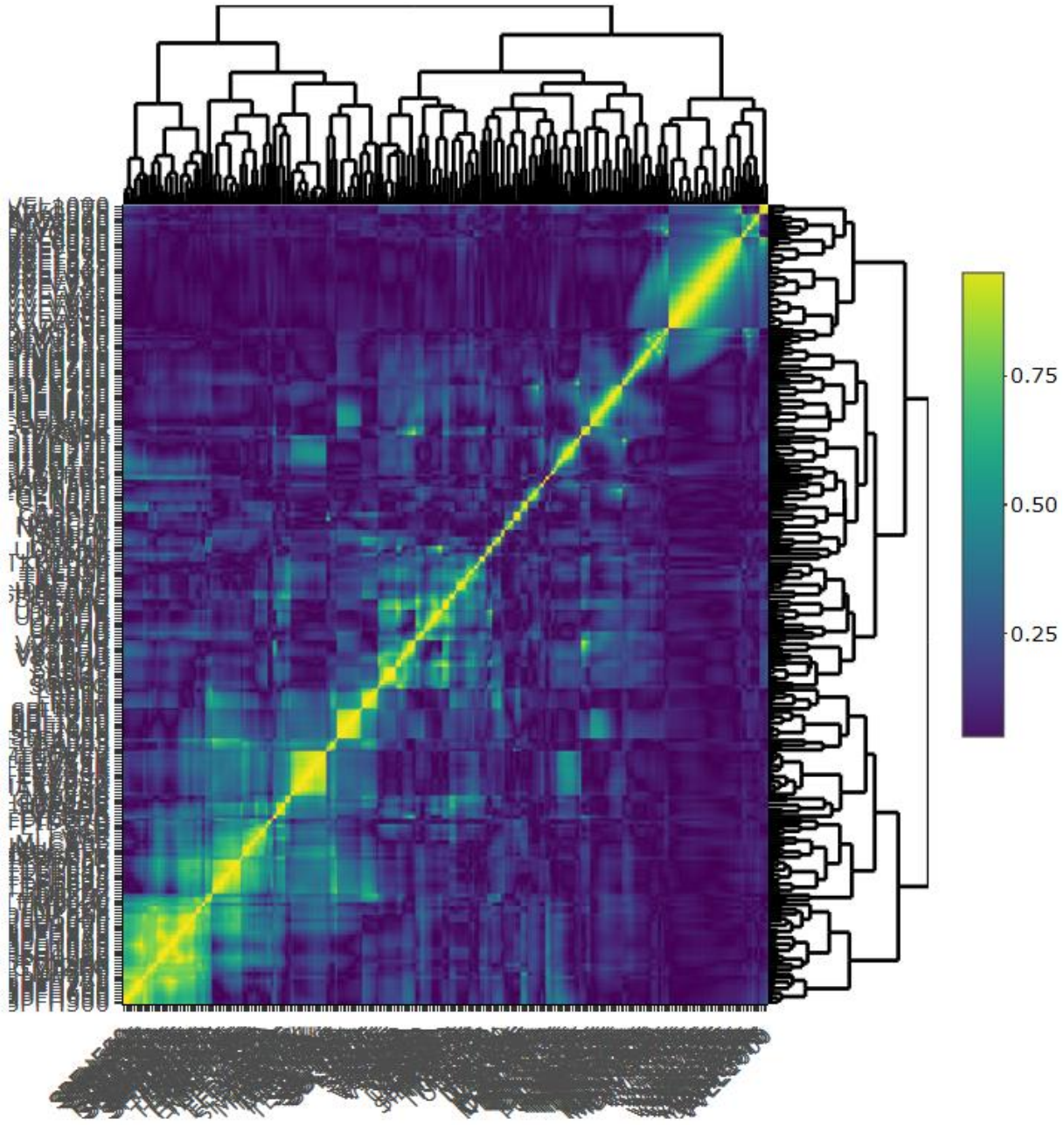


Figure 2.1. Correlation matrix and dendrogram for the maximum value "all" dataset. This correlation matrix shows 253 predictors, which does not include predictors with NAs. The absolute value of predictor vs. predictor correlations is shown in the correlation matrix, where solid yellow represents a correlation of 1.00 (i.e., provides the same information) and solid blue represents a correlation of 0.00. Correlations of 0.4 and higher are considered highly correlated. Cut the dendrogram "branches" at the height which allows for the desired number of clusters. Only one predictor per cluster will be chosen in subsequent steps. Note how there are high correlations outside of the clusters, which is the reasoning for Step 7 in the SWSP.

Table 2.2. Maximum HSS for ten most skillful individual ingredients and combination forecasting indices from Sherburn et al. (2016) study; Table 1.1 from Sherburn (2018). The predictor list was split into two categories, ingredients and combination index. The top ten "ingredients," in order, are: UESHR, U2SHR, HPBLSFC, FGEN725, EFFSHR, FGEN750, W10M, V10M, FGEN700, LR03. The top ten "combination index," in order, are: SHERBE, SHERB35, TKE950, TKE900, TKE925, TKE875, TEVV7, TEVV45K, TEVV3K, TEVV5. Note that Sherburn et al. (2016) removed predictors from this list subjectively, which was based on the relevance of the predictor regarding severe storms found in previous studies.

Ingredient	***Maximum HSS***
U-component of effective bulk shear vector	0.329
U-component of 0-2 km shear vector	0.302
Planetary boundary layer height	0.294
725 hPa frontogenesis	0.292
Effective shear vector magnitude	0.291
750 hPa frontogenesis	0.288
10-m wind magnitude	0.277
V-component of 10-m wind vector	0.275
700 hPa frontogenesis	0.273
0-3 km lapse rate	0.262
Combination Index	***Maximum HSS***
SHERBE	0.374
SHERBS3	0.366
950 hPa TKE	0.347
900 hPa TKE	0.335
925 hPa TKE	0.332
875 hPa TKE	0.284
700 hPa ω * 1000-700 hPa dθ_e/dz	0.257
4.5 km ω * 0-4.5 km dθ_e/dz	0.248
3 km ω * 0-3 km dθ_e/dz	0.247
500 hPa ω * 1000-500 hPa dθ_e/dz	0.246

Table 2.3. Top 50 most skillful predictors determined by the smallest deviance with Y using the NARR maximum value "all" dataset (2063 cases); the dataset only includes the maximum value of each predictor within the sample area (162x162 km^2). Bolded predictors were also in the Sherburn et al. (2016) top skillful predictor list. The top 9 "ingredient" and top 10 "combination index" predictors in the Sherburn et al. (2016) study were within the top 34 predictors (not including MOSH and MOSHE for equivalent comparison). Note that the Sherburn et al. (2016) removed predictors from their top skillful predictor list, including PMSL and PMSL2; therefore, the comparison with the Heidke Skill Score method is more similar than shown. Because this same dataset was used to develop the MOSH and MOSHE, it is no surprise that these two predictors made the top of this list. Three of the four predictors used to calculate the MOSH and MOSHE are also in this list. LR03 is the only missing predictor from this most skillful predictor list, but the minimum value of LR03 is more skillful than its maximum value so this is logical. These results show that the smallest deviance with Y method sufficiently identifies the most skillful predictors, similar to Heidke Skill Score.

Predictor	**Deviance with Y**		**Predictor**	**Deviance with Y**
MOSHE	2332		**HPBLSFC**	2520
MOSH	2341		TEVV8	2522
TKE925	2346		TEVV35K	2525
TKE950	2347		TEVV4K	2528
SHERB3	2411		**U2SHR**	2534
V10M	2423		**TKE875**	2535
SHERB35	2426		TEVV5K	2536
SHERBE	2435		U15SHR	2544
SHERBE35	2438		**FGEN725**	2547
TKE900	2457		**FGEN750**	2548
PMSL2	2459		**FGEN700**	2550
PMSL	2462		U25SHR	2552
UESHR	2462		S15MG	2555
W10M	2477		TEVV55K	2556
TEVV7	2479		FGEN775	2561
MAXTEVV	2485		S2MG	2562
TEVV2K	2488		SBLI800	2563
TEVV5	2489		SBLI750	2564
TEVV25K	2490		SBLI850	2567
EFFSHR	2493		SBLI700	2567
MINTEVV	2496		EFFSRH	2570
TEVV3K	2503		U10M	2570
TKE975	2512		SBLI650	2573
TEVV45K	2513		S25MG	2573
TEVV6K	2518		TKE850	2574

Table 2.4. The first 50 out of 31,878 predictor-predictor correlations using the NARR maximum value "all" dataset (2063 cases). Only the first four decimal places shown for brevity. All predictor-predictor correlations shown are close to 1, which means the two predictors are nearly the same and can be swapped for each other in a prediction model without a significant difference. Correlations show predictor-predictor pairs known to be similar to each other and other pairs which are interesting. For example, the surface-500 hPa lifted index (SBLI500) and maximum lifted index (LIMAX) have nearly identical slopes indicating the maximum lifted index could be reasonably represented by the surface-500 hPa lifted index. The lapse rates between 0 km and 3 km (LR 015, LR02, LR025, LR03) are all similar, which is reasonable since these heights are within the typical boundary layer (PBL heights range from 0.5-4 km in dataset). Also, the updated SHERB (SHERB35) and SHERBE (SHERBE35), which used the 3-5 km lapse rate (LR35) instead of the 700-500 hPa lapse rate (LR75) in the original SHERB (SHERB3) and SHERBE (SHERBE), remained unchanged. Note the correlation between LR35 and LR75 is 0.7421, so swapping these predictors did not make a significant difference in the SHERB(E) composite parameters.

Predictor 1	**Predictor 2**	**Corr**		**Predictor 1**	**Predictor 2**	**Corr**
PMSL	PMSL2	0.9987		VVEL650	VV35	0.9876
SHERBE	SHERBE35	0.9986		VVEL775	VV2	0.9870
SHERB3	SHERB35	0.9954		TEVV45K	TEVV5K	0.9867
SRH2	SRH25	0.9945		TEVV35K	TEVV4K	0.9861
CAPE05	CAPE1	0.9938		VVEL550	VV5	0.9859
SRH25	SRH3	0.9932		SPFH850	SPFH825	0.9858
SBLI500	SBLI550	0.9931		VVEL725	VVEL700	0.9857
SBLI550	SBLI600	0.9912		LR015	LR02	0.9855
SRH15	SRH2	0.9909		LR025	LR03	0.9854
VVEL725	VV25	0.9908		VVEL500	VV5	0.9848
SBLI700	SBLI750	0.9903		SPFH875	SPFH850	0.9847
SPFH725	SPFH700	0.9900		VVEL750	VVEL725	0.9845
VVEL600	VV4	0.9898		TEVV25K	MAXTEVV	0.9841
SBLI650	SBLI700	0.9896		VVEL700	VV3	0.9838
SPFH800	SPFH775	0.9893		TEVV2K	MAXTEVV	0.9829
SPFH750	SPFH725	0.9891		VVEL775	VVEL750	0.9825
SBLI600	SBLI650	0.9890		VVEL650	VV3	0.9820
VVEL550	VV45	0.9889		SBLI500	LIMAX	0.9819
SBLI750	SBLI800	0.9886		SBLI550	LIMAX	0.9818
SPFH775	SPFH750	0.9884		TEVV5K	TEVV6K	0.9818
TEVV4K	TEVV45K	0.9881		SPFH1512	SPFH1815	0.9816
SPFH825	SPFH800	0.9879		TEVV25K	TEVV3K	0.9813
LR02	LR025	0.9879		VVEL600	VV45	0.9812
VVEL500	VV55	0.9876		SPFH900	SPFH875	0.9808
TEVV3K	TEVV35K	0.9876		TEVV5K	TEVV55K	0.9806

Table 2.5. The first 836-886 out of 31,878 predictor-predictor correlations using the NARR maximum value "all" dataset (2063 cases). Only the first four decimal places shown for brevity. Table showing correlations around 0.75. Correlations around 0.75 are still showing predictor-predictor pairs known to be similar to each other; for example, DIVG800 and DIVG750. Statistically speaking, a 0.75 correlation is a strong linear relationship, and this can be seen in these data. For example, MUCAPE has a linear relationship with VGP, which is a function of CAPE. The temperature lapse rate calculated in the 700-500 hPa level (LR75; around 3-5.5 km) has an inverse relationship with the temperature lapse rate calculated in the 3-6 km level (LR36), but this is a mistake in the sign of LR75 (all positive values) in the dataset. Note the correlation between LR35 and LR36 is 0.9398.

Predictor 1	**Predictor 2**	**Corr**		**Predictor 1**	**Predictor 2**	**Corr**
SRH1	S15MG	0.7572		TKE925	W10M	0.7511
SPFH925	TMPLCL	0.7572		SPFH725	SPFH1290	0.7510
TMP925	SPFH9060	0.7572		SPFH950	SPFH750	0.7501
SPFH700	SPFH1512	0.7571		SRH15	S15MG	0.7501
SPFH825	SPFH30T0	0.7569		SRH05	V05SHR	0.7501
TMP850	TMP500	0.7567		V1SHR	S1MG	0.7500
SPFH975	SPFH775	0.7565		DIVG950	DIVG925	0.7499
VVEL775	VVEL650	0.7563		TMP1000	TEDF1050	0.7499
SPFH650	TMP700	0.7562		SPFH800	SPFH30T0	0.7497
LR75	LR36	-0.7562		TMPLFC	DPT2M	0.7490
TMP925	TMP700	0.7560		SPFH1000	TMP1000	0.7486
V15SHR	V05SHR	0.7546		SPFH750	TMP850	0.7483
MUCAPE	TEDF06	0.7544		TEVV5	TEVV8	0.7483
DIVG800	DIVG750	0.7544		EFFSRH	SCP	0.7482
VVEL700	VVEL550	0.7544		SPFH775	SPFH550	0.7481
TEVV8	TEVV5K	0.7541		TMP925	SPFH30T0	0.7479
TEDF02	TEDF06	0.7541		TEDF02	TEDF05	0.7475
SPFH650	HGT0DEG	0.7540		TEVV8	TEVV55K	0.7475
TMPLCL	HGT0DEG	0.7527		VV2	VV35	0.7470
FGEN975	FGEN900	0.7526		TMP1000	TEDF06	0.7469
SPFH1512	DPT2M	0.7526		S25MG	S1MG	0.7469
MUCAPE	VGP	0.7517		VVEL800	VV3	0.7465
DIVG700	DIVG650	0.7516		VVEL650	VVEL500	0.7456
DIVG750	DIVG700	0.7514		SPFH900	DPT2M	0.7450
S5MG	S3MG	0.7513		TMP1000	DPT2M	0.7448

Table 2.6. The first 2196-2246 out of 31,878 predictor-predictor correlations using the NARR maximum value "all" dataset (2063 cases). Only the first four decimal places shown for brevity. Table showing correlations around 0.50, which is the cutoff for deciding highly correlated predictors. Not all atmospheric variables have linear relationships, which means their correlation could be around 0.5 while having a significant relationship between the predictors. This can be seen in these data; for example, the 0-0.5 km temperature lapse rate (LR005) and the surface temperature (TMP2M) have a correlation of 0.5011. The surface temperature influences the lapse rate at the surface, and the lapse rate is calculated as the vertical temperature change with height, so the relationship is not linear. Because related predictors are showing correlations around 0.4 and higher, it may be beneficial to only include predictors in the model with predictor-predictor correlations below 0.4, which was done with the Chapter II training datasets. For example, 0-0.5 km CAPE and 0-0.5 km lapse rate have a correlation of -0.4034.

Predictor 1	**Predictor 2**	**Corr**		**Predictor 1**	**Predictor 2**	**Corr**
TEVV5K	MOSHE	0.5028		DIVG275	TEVV6K	0.4999
TEVV45K	MOSHE	0.5028		TKE950	PMSL2	-0.4999
S5MG	S15MG	0.5028		DIVG750	MDIV1815	0.4999
MDIV6030	VVEL1290	0.5028		DIVG925	VVEL9060	0.4998
SPFH1000	MUCAPE	0.5024		DIVG250	MINTEVV	0.4997
CAPE3	SHERBE35	0.5024		SPFH725	TMP1000	0.4997
EFFSDPTH	TEVV6K	0.5022		TKE950	TKE850	0.4996
DIVG900	VV25	0.5020		SRH25	U2SHR	0.4995
U25SHR	U05SHR	0.5018		TEVV2K	TEVVDF	0.4994
FGEN925	FGEN800	0.5018		SBLI650	LR025	-0.4994
FGEN725	FGEN600	0.5017		TEVV35K	MOSH	0.4997
CAPE25	TEDF06	0.5014		TEDF1085	TEVV8	0.4990
LR005	TMP2M	0.5011		FGEN900	SBLI650	0.4988
TEDF05	SHERBE35	0.5011		TEDF03	LR02	0.4986
V3SHR	V05SHR	0.5011		VVEL725	DIVG825	0.4986
SPFH975	MUCAPE	0.5010		DIVG250	TEVV7	0.4985
SPFH925	MLCAPE	0.5010		FGEN975	SBLI750	0.4983
NSBCAPE	TEVV3K	0.5008		SBLI800	TEDF1085	-0.4982
U2SHR	S05MG	0.5008		SPFH500	TMP925	0.4982
V2SHR	S1MG	0.5006		SRH05	U15SHR	0.4982
TEDF05	SHERBE	0.5005		SPFH6030	TEDF025	0.4982
MUCAPE	TEVV5K	0.5005		TEVV45K	MOSH	0.4982
TEVV5	MOSH	0.5001		NMLCAPE	TEVV6K	0.4982
STP	SHERBE35	0.5000		TKE950	PMSL	-0.4981
UESHR	SHERB35	0.5000		CBSS	SHERBE35	0.4981

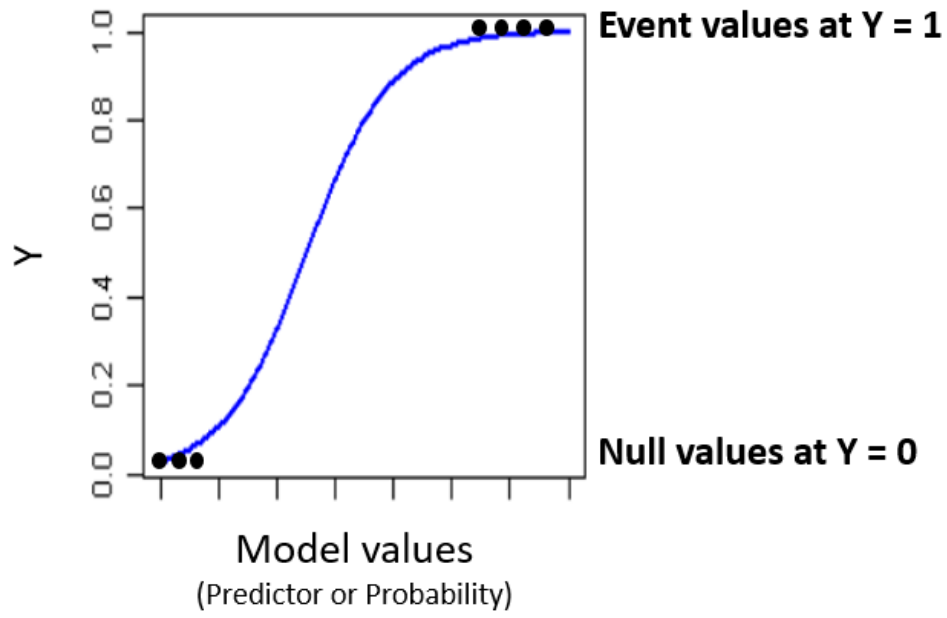


Figure 2.2. Graphic showing an example of an ideal probabilistic model, which has an s-shaped curve (blue line). Theoretical model values are marked with black circles. On the y-axis, Y, is the actual (observed) probability, that is, 0.0 for null cases and 1.0 for event cases; therefore, all Y values (black circles) are plotted along Y=0 or Y=1. On the x-axis, the model values are shown; typically, these are the predicted probability values calculated by the probabilistic model, but they can alternatively be the values of a predictor to visualize the predictor's relationship with Y. The more separation between the 0 and 1 model values, the closer model is to an "s-shaped curve," and therefore, the more skillful the model (or predictor).

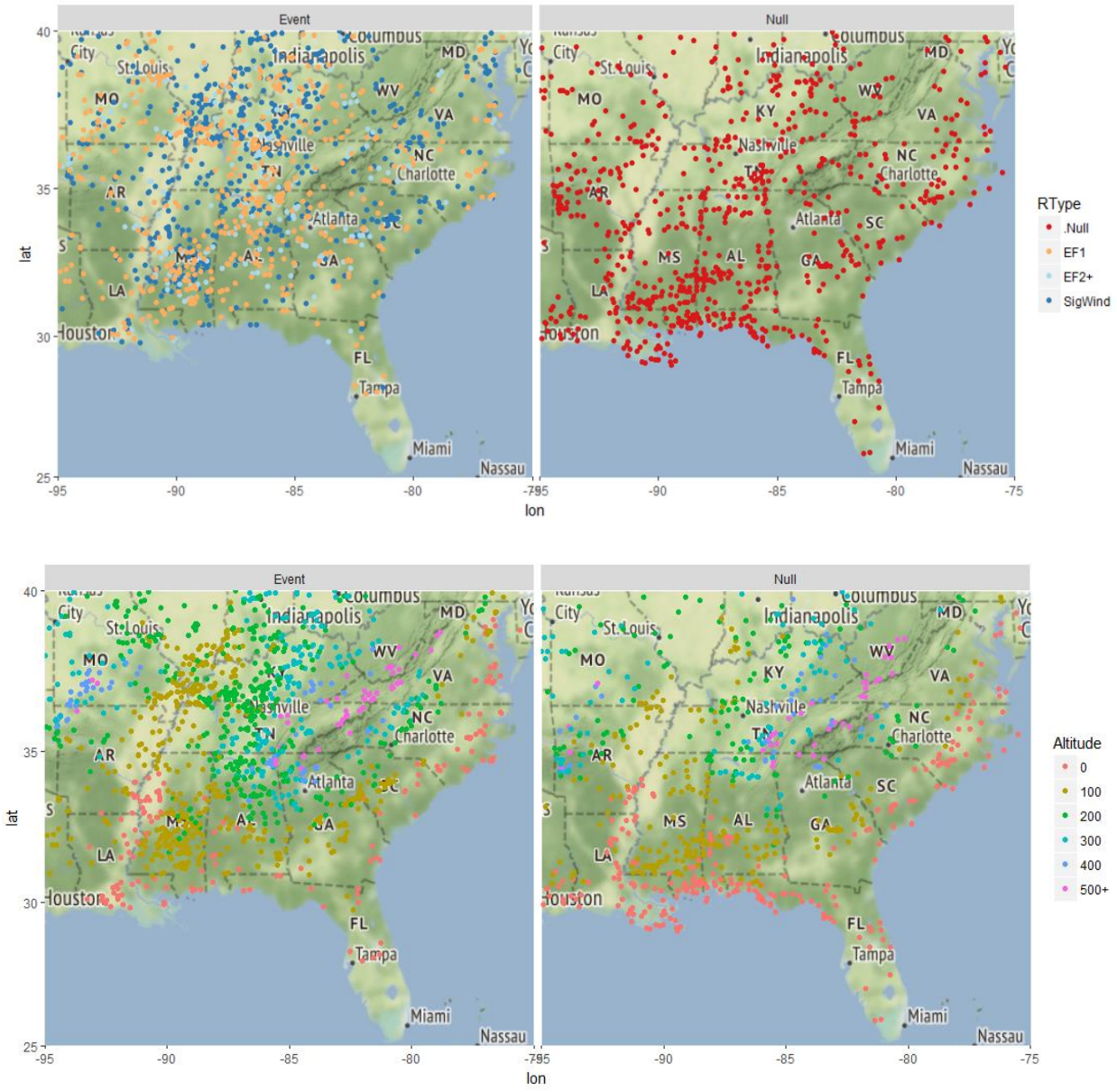


Figure 2.3. Geographical maps showing the locations of all cases in the "all" dataset (2063 cases). Event cases, NWS official storm reports of tornadoes and severe wind speeds, are shown on the left. Null cases, NWS issued warning but no storm reports, are shown on the right. Note some dots are covered by others. Google Maps background. Top map: Red dots show the central location of false alarms (null cases). Orange dots are the location of EF1 tornadoes, light blue dots are EF2 and stronger tornadoes (i.e., significant tornadoes), and dark blue dots are severe wind speeds (i.e., wind gust ≥ 65 knots). Bottom map: Colors depict altitude in meters (surface elevation of storm report or null). Altitudes for each location provided by USGS National Elevation Dataset (Schneider 2016). Most null cases were near the coastline and/or at lower elevations, which could be due to observation bias (i.e., more populated areas).

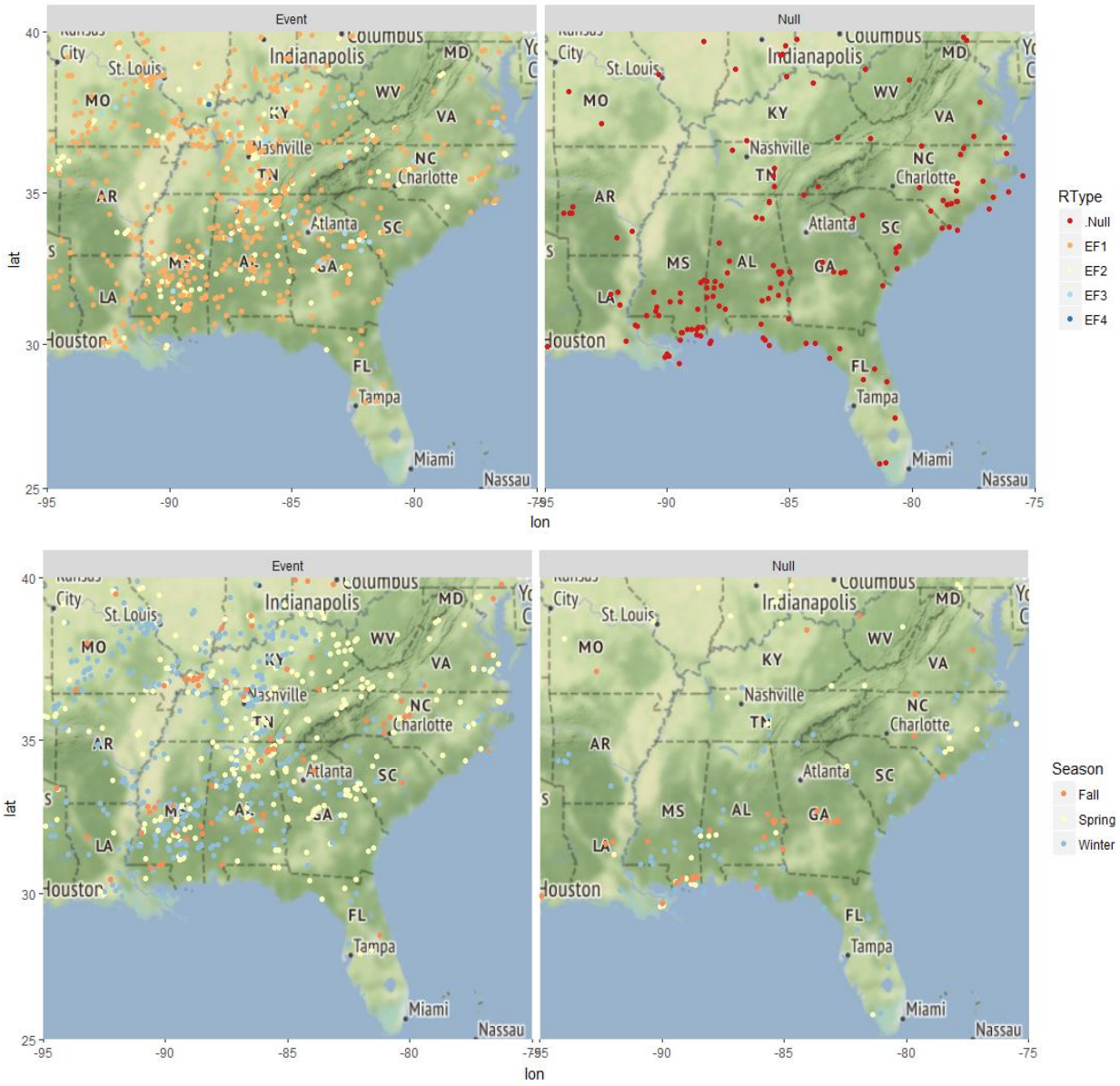


Figure 2.4. Geographical maps showing the locations of all cases in the "tornado" dataset (865 cases) and the severity of each tornado case. Event cases, NWS official storm reports of tornadoes, are shown on the left. Null cases, NWS issued tornado warning but no official tornado reports, are shown on the right. Note some dots are covered by others. Tornadoes associated with tropical cyclone landfall are not included in the dataset. Google Maps background. Top map: Red dots show the central location of NWS false alarms (145 null cases). Orange dots are the location of EF1 tornadoes (519 cases), yellow dots are EF2 tornadoes (156 cases), light blue dots are EF3 tornadoes (41 cases), and dark blue dots are EF4 tornadoes (4 cases). Bottom map: Orange dots show cases that occurred in Fall, yellow dots occurred in Spring, and blue dots occurred in Winter. Majority of cases occurred during meteorological Winter and Spring seasons, which is also true for severe wind cases.

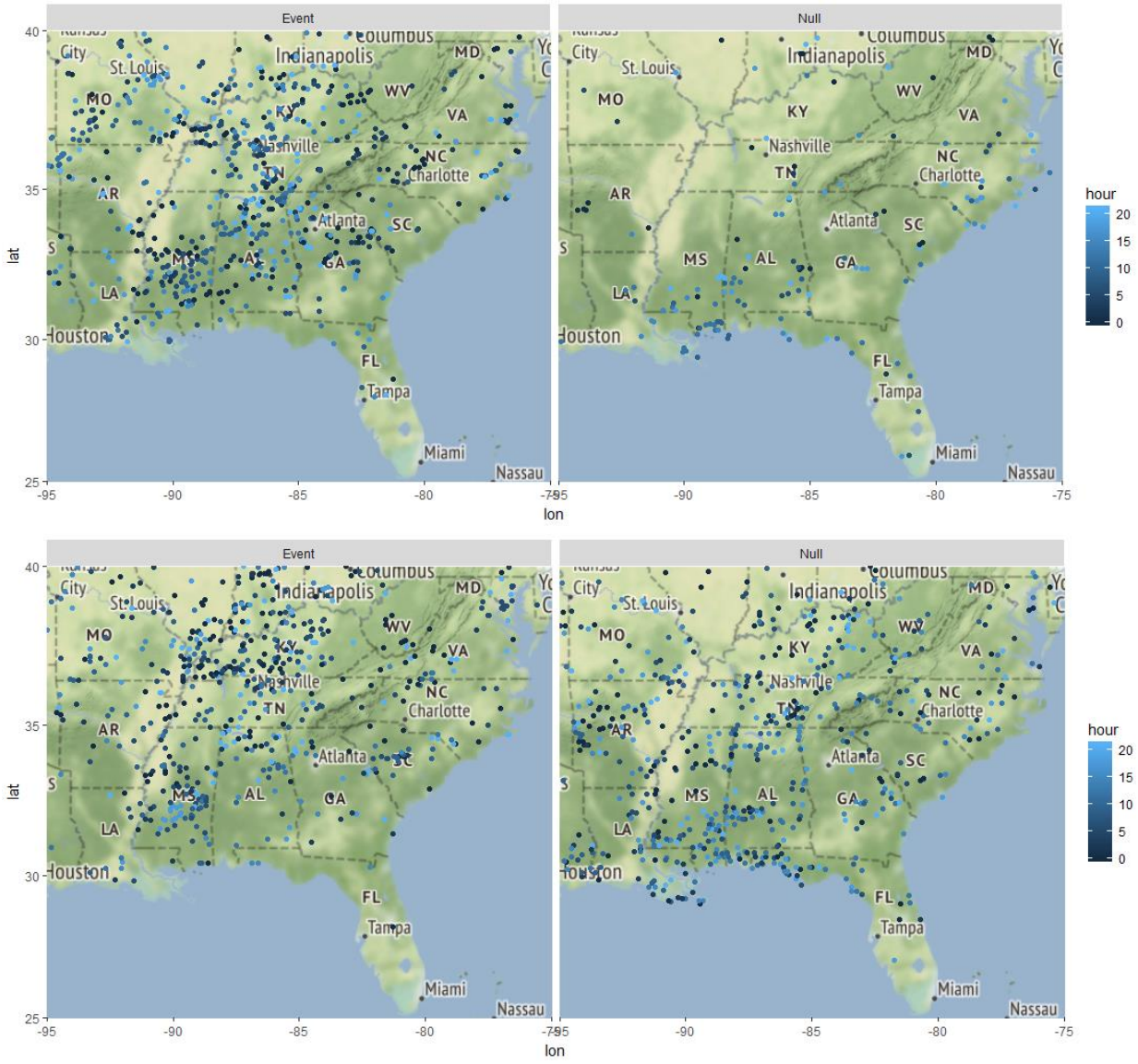


Figure 2.5. Geographical maps showing the locations of cases and time of day for each case. Event cases, NWS official storm reports, are shown on the left. Null cases, NWS issued warning but no official storm reports, are shown on the right. Note some dots are covered by others. Google Maps background. Top map: Color scale shows the hour (UTC) the case occurred for all cases in the "tornado" dataset (720 events and 145 nulls). Tornadoes associated with tropical cyclone landfall are not included in the dataset. Bottom map: Color scale shows the hour (UTC) the case occurred for all cases in the "severe wind speed" dataset (605 events and 593 nulls). In both datasets, most cases occurred in the evening or overnight hours (around 0-12 UTC which are depicted with darker shades of blue).

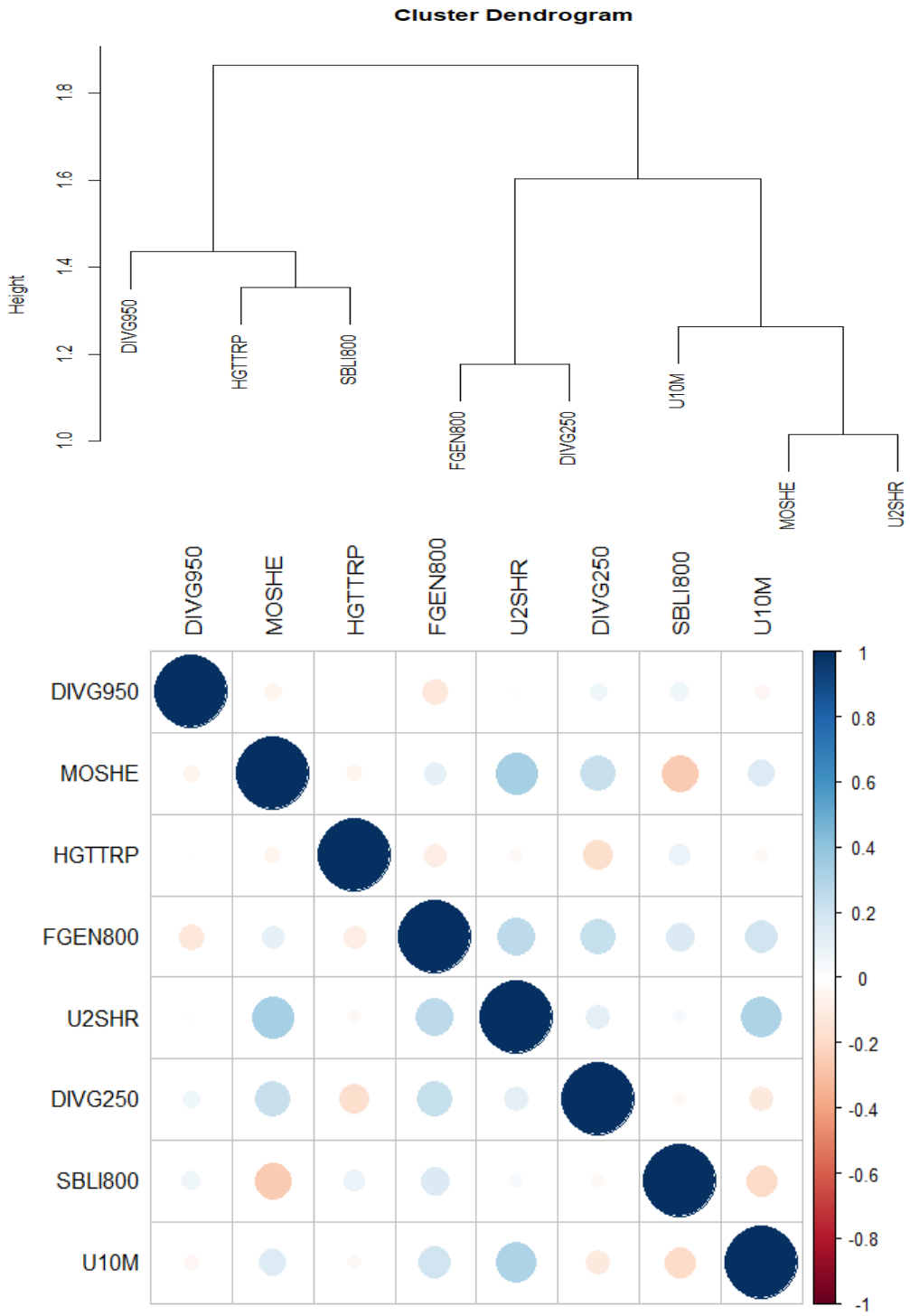


Figure 2.6. Graphics showing the relationship between predictors in the final probabilistic model produced using the "all" dataset with all available variables (i.e., includes severe weather composite parameters). Top graph: Dendrogram groups together predictors that are most alike based on the predictor-predictor correlations. Bottom graph: Correlation matrix depicts higher correlations with a larger circle. Blue colors are positive correlations (direct relationship) and red colors are negative correlations (inverse relationship). In this final cut, all correlations are below 0.4. These graphics take into account the sign of the correlation, which is not used in the SWSP.

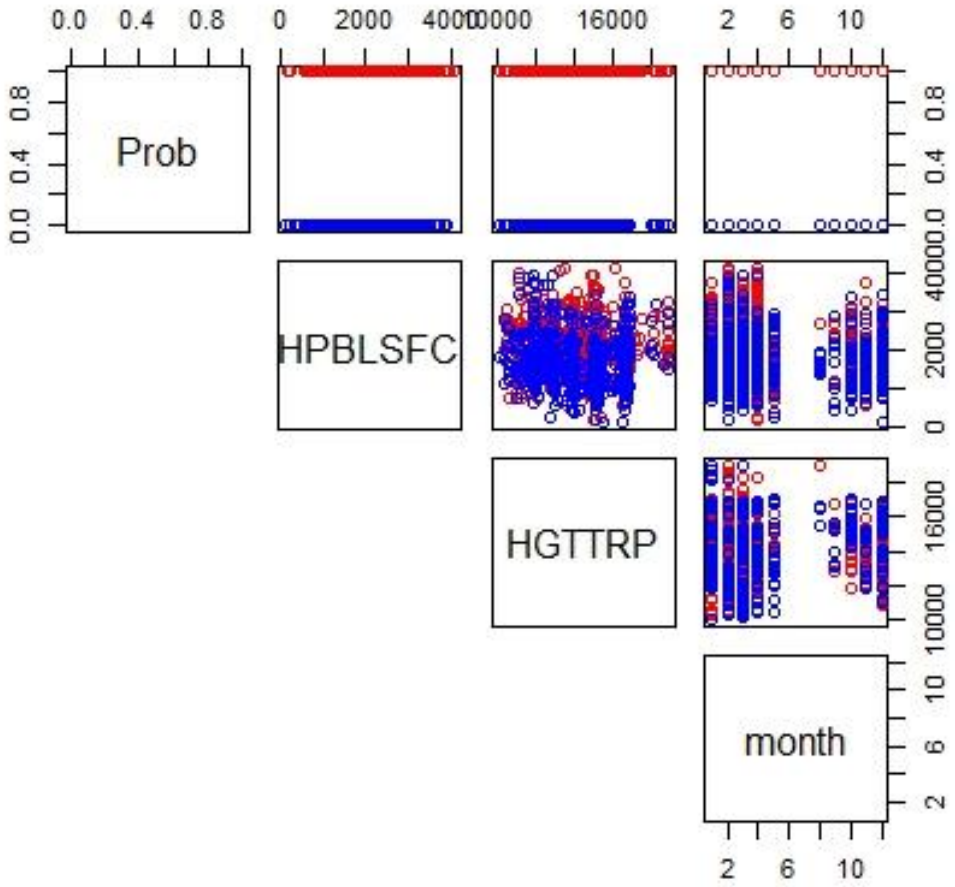


Figure 2.7. Comparing predictor values in the maximum value "all" dataset (2063 cases) for HGTTRP and HPBLSFC. Prob is Y. Red circles represent event cases (Y=1) and blue circles represent null cases (Y=0). For all cases, tropopause height (HGTTRP) ranged from 10.1 km to 18.9 km, and planetary boundary layer height (HPBLSFC) ranged from 105 m to 4.1 km. HGTTRP values had a more limited range as well as higher minimums in the Fall and early Winter months (September-December) compared to late Winter and Spring months (January-May). Many null cases had planetary boundary layer heights in the 1 km to 3 km range and many event cases had planetary boundary layer heights in the 2 km to 4 km range, but both null and event cases existed throughout the entire range. The minimum value of HPBLSFC was a significant predictor (one of top 15 predictors in minimum dataset). HGTTRP was a significant predictor in both the maximum and minimum datasets.

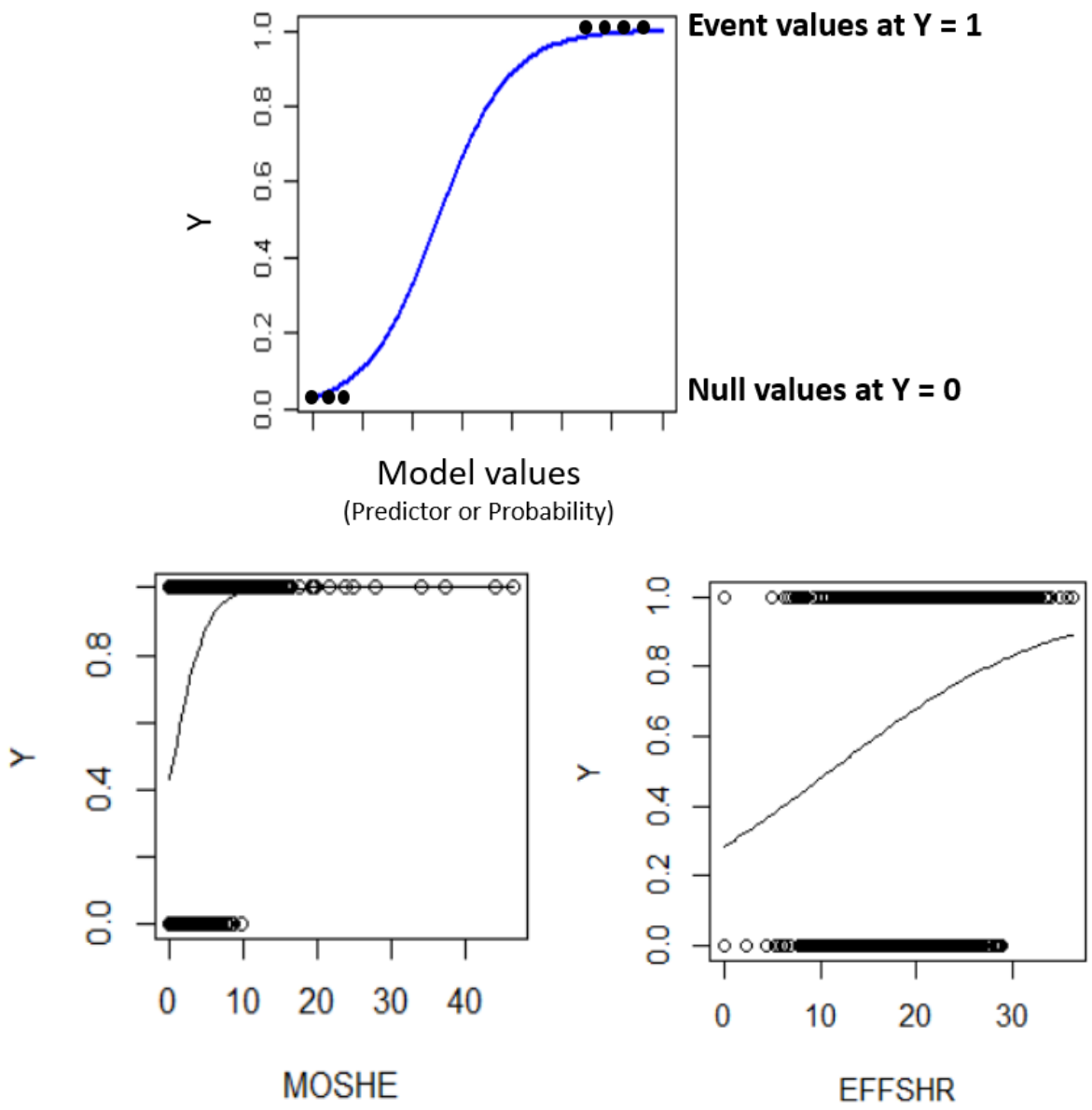


Figure 2.8. Graphics showing probabilistic models using the maximum value dataset. Top graph is an example of an ideal probabilistic model, which has an s-shaped curve (blue line) and theoretical model values marked with black circles. On the y-axis, Y, is the actual (observed) probability, 0.0 for null cases and 1.0 for event cases; therefore, all Y values (circles) are plotted along Y=0 or Y=1. On the x-axis, the model values are shown, which are predictor values in the bottom graphs. The more separation between the 0 and 1 model values, the closer the model is to an "s-shaped curve," and therefore, the more skillful the model. The bottom graphs show the distribution of the predictor values, modified SHERBE (MOSHE) and effective bulk shear magnitude (EFFSHR), with respect to Y. For reference, the MOSHE value above 9.8 has a clear separation between events and nulls for 71 out of 1326 event cases (5%), and EFFSHR value above 29 has a clear separation between events and nulls for 117 out of 1326 event cases (9%); this value alone is not an indicator for model skill.

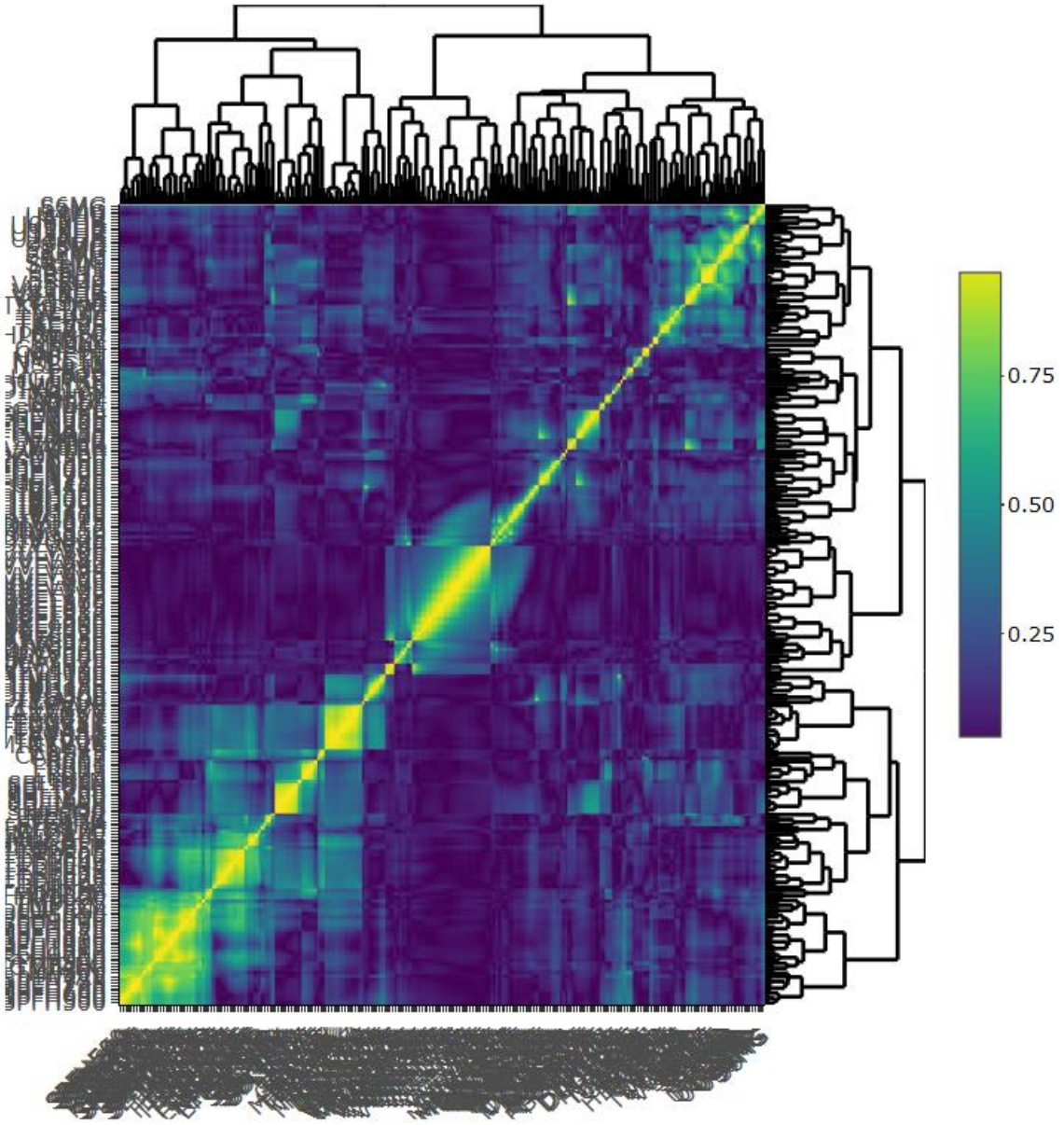


Figure 2.9. Correlation matrix and dendrogram for the maximum value "all" dataset. This correlation matrix shows 242 predictors, which does not include predictors with NAs or severe weather composite parameters. The absolute value of predictor vs. predictor correlations is shown in the correlation matrix, where solid yellow represents a correlation of 1.00 (i.e., provides the same information) and solid blue represents a correlation of 0.00. Correlations of 0.4 and higher are considered highly correlated. Cut the dendrogram "branches" at the height which allows for the desired number of clusters. Only one predictor per cluster will be chosen in subsequent steps. Note how there are high correlations outside of the clusters, which is the reasoning for Step 7 in the SWSP.

Table 2.7. Top 50 most skillful predictors determined by the smallest deviance with Y using the NARR minimum value "all" dataset (2063 cases); the dataset only includes the minimum value of each predictor within the sample area (162x162 km^2). The only missing predictor in Table 2.3 from the top skillful predictors in Sherburn et al. (2016) was 0-3 km temperature lapse rate (LR03), which is shown near the top of the table. The minimum value of LR03 is more skillful than its maximum value so this is logical. The minimum value of the mean sea level pressure (PMSL and PMSL2) is most likely showing skill because of the association between stronger surface low pressure systems (lower PMSL) and increased chance of severe weather.

Predictor	**Deviance with Y**		**Predictor**	**Deviance with Y**
PMSL2	2422		VV35	2576
PMSL	2425		SBLI700	2576
SHERB35	2453		VV5	2577
SHERB3	2453		VVEL500	2577
HPBLSFC	2469		VV55	2579
TEDF1085	2509		VVEL550	2579
TEDF1070	2527		VV4	2579
TKE950	2528		DIVG800	2580
W10M	2530		LR015	2580
DIVG875	2537		VVEL1815	2581
MOSH	2547		VV45	2581
TEDF02	2549		VVEL600	2582
TEDF025	2553		VV25	2582
DIVG900	2553		TEDF04	2583
DIVG850	2554		S15MG	2583
TKE925	2555		VV3	2584
DIVG825	2560		VV2	2585
LR025	2561		VVEL1512	2586
TEDF1050	2562		TEDF045	2586
LR03	2563		VVEL650	2586
TEDF03	2566		DIVG700	2588
SBLI650	2568		S2MG	2588
LR02	2568		VVEL700	2589
TEDF035	2571		DIVG725	2589
VV6	2574		DIVG650	2589

Table 2.8. Difference between the deviance with Y of the maximum value and minimum value for each predictor, and labeling of most skillful value (i.e., maximum or minimum) for each predictor. The deviance with Y was calculated for each predictor in the maximum value dataset and minimum value dataset, separately. For each predictor in the predictor list, the deviance with Y with the maximum value of the predictor was subtracted from the deviance with Y with the minimum value of the predictor, which gave the deviance with Y difference between the two predictors (Dev Diff). Because the smallest deviance with Y indicates a more skillful predictor, if this deviance difference was negative, the minimum value of the predictor was more skillful than the maximum value of the predictor. The predictors that were more skillful (closer relationship with Y) when using its minimum value in the sample area (162x162 km^2) are shaded blue; 117 of the 251 predictors are more skillful at their minimum value. This means that the predictor has a greater difference between an event case and a null case when comparing the minimum values of the predictor rather than the maximum values; the larger the deviance difference (absolute value), the larger the difference in skill between the maximum value predictor and the minimum value predictor. Because the maximum or minimum value was taken over a large sample area, some predictors may not be skillful near the location of the severe weather hazard, but they are skillful when evaluating the 162x162 km^2 area surrounding the event location. Regarding the large area encompassing a convective environment, these results appear to make physical sense.

Predictor Name	Dev Diff	Predictor Name	Dev Diff	Predictor Name	Dev Diff	Predictor Name	Dev Diff
TEDF03	-76	U4SHR	41	NVWSTRP	51	TEVV5	167
TEDF035	-90	U3SHR	51	TMP2M	-24	TEVV7	173
TEDF04	-79	U25SHR	64	DPT2M	-17	TEVV8	111
TEDF045	-75	U2SHR	66	RH2M	7	TEVV2K	165
TEDF05	-86	U15SHR	58	U10M	50	TEVV25K	167
TEDF055	-80	U1SHR	7	V10M	176	TEVV3K	158
TEDF06	-64	U05SHR	-30	THSFC	-43	TEVV35K	138
STP	64	V6SHR	19	PMSL	-37	TEVV4K	139
SCP	72	V5SHR	4	PMSL2	-37	TEVV45K	156
EHI	58	V4SHR	12	HGT0DEG	-5	TEVV5K	134
CBSS	25	V3SHR	11	HPBLSFC	-52	TEVV55K	111
VGP	83	V25SHR	17	HGTTRP	-28	TEVV6K	152
SHERB3	43	V2SHR	25	MUPL	-65	MAXTEVV	175
SHERBE	224	V15SHR	33	RHHL1	7	MINTEVV	161
LR005	-23	V1SHR	45	PWAT	-6	TEVVDF	-16
LR01	-32	V05SHR	49	CLTPRS	1	W10M	53
LR015	-41	S6MG	4	VV2	-105	MOSH	206
LR02	-44	S5MG	-5	VV25	-108	MOSHE	345
LR025	-39	S4MG	8	VV3	-106	SHERB35	26
LR03	-36	S3MG	15	VV35	-114	SHERBE35	220
LR75	0	S25MG	35	VV4	-111		
LR36	-8	S2MG	26	VV45	-109		
LR35	-5	S15MG	29	VV5	-112		
U6SHR	-7	S1MG	31	VV55	-111		
U5SHR	19	S05MG	17	VV6	-115		

Table 2.8. (continued).

Predictor Name	Dev Diff	Predictor Name	Dev Diff	Predictor Name	Dev Diff	Predictor Name	Dev Diff
SPFH1000	-58	TKE1000	84	FGEN975	-15	SRH3	5
SPFH975	-45	TKE975	167	FGEN925	-7	CAPE05	5
SPFH950	-29	TKE950	182	FGEN900	0	CAPE1	6
SPFH925	-14	TKE925	210	FGEN875	21	CAPE15	2
SPFH900	0	TKE900	155	FGEN850	54	CAPE2	4
SPFH875	5	TKE875	126	FGEN825	90	CAPE25	27
SPFH850	6	TKE850	102	FGEN800	112	CAPE3	56
SPFH825	5	DIVG1000	-49	FGEN775	126	NMLCAPE	4
SPFH800	4	DIVG975	-38	FGEN750	133	NSBCAPE	-11
SPFH775	-2	DIVG950	-47	FGEN725	131	MLCAPE	6
SPFH750	-11	DIVG925	-89	FGEN700	129	MUCAPE	3
SPFH725	-14	DIVG900	-127	FGEN650	92	NMLCIN	-13
SPFH700	-12	DIVG875	-146	FGEN600	104	NSBCIN	-2
SPFH650	-15	DIVG850	-135	FGEN550	72	MUCIN	-18
SPFH600	-5	DIVG825	-130	FGEN500	19	SBLI500	26
SPFH550	2	DIVG800	-109	MDIV850	-78	SBLI550	18
SPFH500	4	DIVG775	-92	MDIV30T0	-34	SBLI600	10
TMP1000	-53	DIVG750	-90	MDIV6030	-36	SBLI650	-5
TMP925	-11	DIVG725	-98	MDIV9060	-55	SBLI700	9
TMP850	-2	DIVG700	-99	MDIV1290	-46	SBLI750	35
TMP700	-15	DIVG650	-97	MDIV1512	-31	SBLI800	53
TMP500	4	DIVG600	-69	MDIV1815	-28	LIMIN	7
VVEL1000	-4	DIVG550	-50	SPFH30T0	-26	LIMAX	25
VVEL975	6	DIVG500	-3	SPFH6030	-6	MULFC	10
VVEL950	0	DIVG450	17	SPFH9060	13	TMPLFC	-12
VVEL925	-7	DIVG400	20	SPFH1290	18	SBLCL	4
VVEL900	-30	DIVG350	65	SPFH1512	14	TMPLCL	0
VVEL875	-57	DIVG300	110	SPFH1815	11	EFFSDPTH	17
VVEL850	-75	DIVG275	107	VVEL30T0	-31	EFFIDPTH	1
VVEL825	-85	DIVG250	113	VVEL6030	-54	EFFSHR	169
VVEL800	-90	DIVG225	103	VVEL9060	-65	UESHR	191
VVEL775	-95	DIVG200	105	VVEL1290	-89	VESHR	90
VVEL750	-98	DIVG175	55	VVEL1512	-104	EFFSRH	112
VVEL725	-100	DIVG150	4	VVEL1815	-109	TEDF1050	-88
VVEL700	-101	DIVG125	-15	SRH05	11	TEDF1070	-88
VVEL650	-104	DIVG100	-18	SRH1	9	TEDF1085	-109
VVEL600	-108	DZDX500	-26	SRH15	9	TEDF8570	-1
VVEL550	-111	DZDY500	40	SRH2	8	TEDF02	-53
VVEL500	-112	FGEN1000	-37	SRH25	7	TEDF025	-63

Table 2.9. Top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR maximum and minimum value "all" dataset (2063 cases); see Table 2.8. The value of each predictor was taken within a sample area (162x162 km^2). The **minimum value** predictors are bolded. Because this same dataset was used to develop the MOSH and MOSHE, it is no surprise that these two predictors made the top of this list. All predictors in the predictor list that are HSLC severe weather composite parameters (which all used the same or similar case list as this study) are among the most skillful predictors, which brings more evidence that the smallest deviance with Y method compares well with other methods.

Predictor	**Deviance with Y**		**Predictor**	**Deviance with Y**
MOSHE	2332		TEVV45K	2513
MOSH	2341		TEVV6K	2518
TKE925	2346		TEVV8	2522
TKE950	2347		TEVV35K	2525
SHERB3	2411		**TEDF1070**	2527
PMSL2	2422		TEVV4K	2528
V10M	2423		U2SHR	2534
PMSL	2425		TKE875	2535
SHERB35	2426		TEVV5K	2536
SHERBE	2435		**DIVG875**	2537
SHERBE35	2438		U15SHR	2544
TKE900	2457		FGEN725	2547
UESHR	2462		FGEN750	2549
HPBLSFC	2469		**TEDF02**	2549
W10M	2477		FGEN700	2550
TEVV7	2479		U25SHR	2552
MAXTEVV	2485		**TEDF025**	2553
TEVV2K	2488		**DIVG900**	2553
TEVV5	2489		**DIVG850**	2554
TEVV25K	2490		S15MG	2555
EFFSHR	2493		TEVV55K	2556
MINTEVV	2496		**DIVG825**	2560
TEVV3K	2503		FGEN775	2561
TEDF1085	2509		**LR025**	2561
TKE975	2512		S2MG	2562

Table 2.10. Evaluation of a probabilistic model's ability to give an interpretable probability. The table shows the evaluation for the probabilistic model trained using the combined NARR maximum and minimum value "all" dataset shown in section 2.6.4. For example, the model predicted a probability between 90% to 100% for 391 cases and 368 of those cases were an actual event case (Y=1), which equates to 94.12%. The actual percentage of event cases was 94.12%, which is the actual probability. If the model predicted perfectly, the average actual probability for the 90% to 100% range is 95%; therefore, the model performed very well for that prediction range. Note that this is not related to the optimal cutoff (probability threshold) for the model; that threshold just determines whether to mark a prediction as a null or an event during model evaluation for model comparisons. These results show proof that the skillful probabilistic models give an interpretable probability. An interpretable probability means that if the model gives a prediction of 50%, an event will occur, on average, 50% of the time if there were an infinite number of cases. For reference, there were 1325 event cases and 738 null cases in the training dataset (2063 total cases). Also, the 95% confidence interval of the actual probability for this model was +/- 8%; looking at this table, the most uncertainty in the probability came from the lower predictions. Predicted probabilities of 0.5-1 (50%-100%) had closer to [-1% to +4%] uncertainty. There were 1325 event cases and 738 null cases in the training dataset, so increasing the number of null cases may help decrease the model uncertainty.

Predicted Probability (%)	**Actual Event Cases (Y = 1)**	**Total Cases (Events & Nulls)**	**Actual Probability (%)**
90-100	368	391	94.12
80-89.99	304	355	85.63
70-79.99	228	287	79.44
60-69.99	162	254	63.78
50-59.99	93	171	54.39
40-49.99	64	168	38.10
30-39.99	40	128	31.25
20-29.99	45	149	30.20
10-19.99	12	103	11.65
0-9.99	9	57	15.79

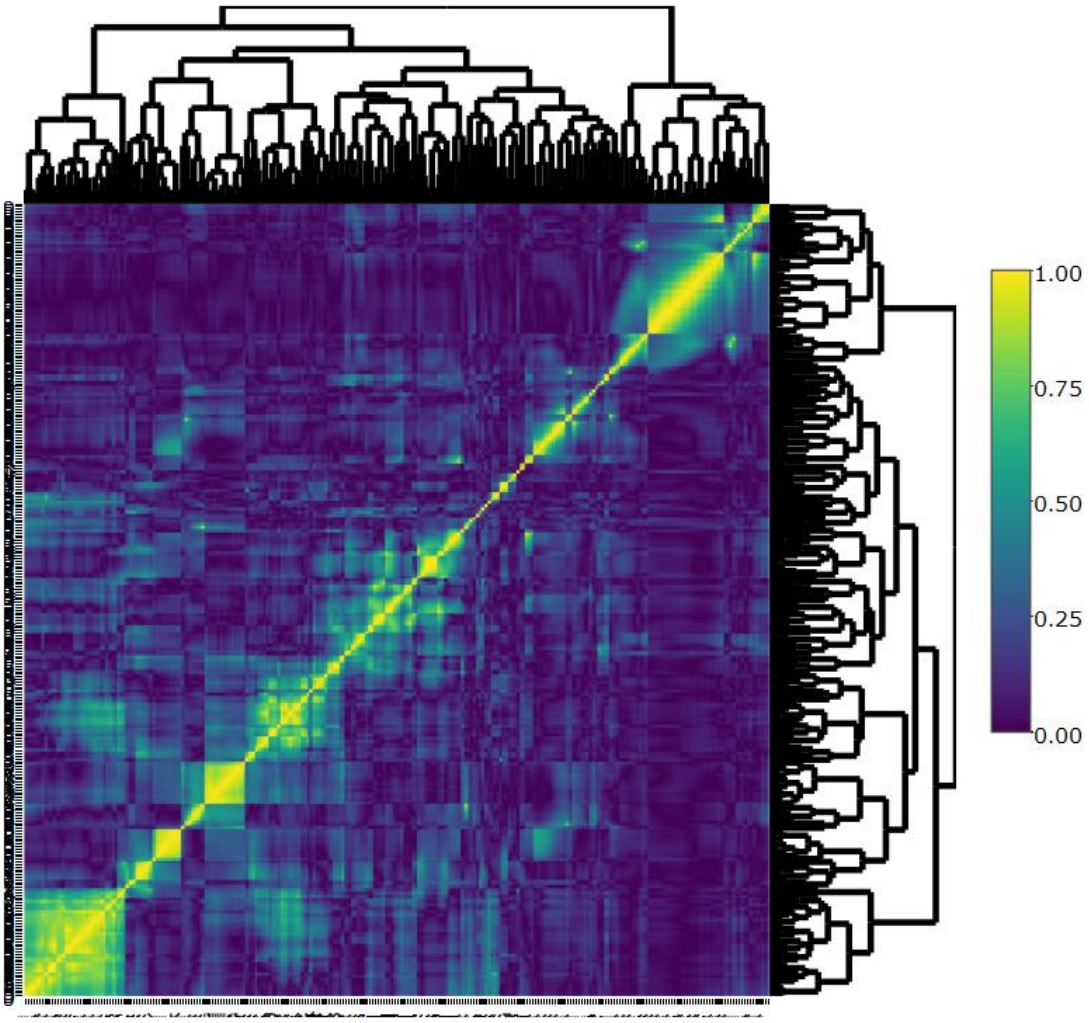


Figure 2.10. Correlation matrix and dendrogram for the maximum value "tornado" dataset. The absolute value of predictor vs. predictor correlations is shown in the correlation matrix, where solid yellow represents a correlation of 1.00 (i.e., provides the same information) and solid blue represents a correlation of 0.00. Correlations of 0.4 and higher are considered highly correlated. Cut the dendrogram "branches" at the height which allows for the desired number of clusters. Only one predictor per cluster will be chosen in subsequent steps. Note how there are high correlations outside of the clusters, which is the reasoning for Step 7 in the SWSP.

Table 2.11. Top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR maximum and minimum value "tornado" dataset (865 cases). The value of each predictor was taken within a sample area (162x162 km^2). The **minimum value** predictors are bolded. Because this same dataset was used to develop the MOSH and MOSHE, it is no surprise that these two predictors made the top of this list. All predictors in the predictor list that are HSLC severe weather composite parameters (which all used the same or similar case list as this study) are among the most skillful predictors. However, all versions of the MOSH and SHERB predictors either did not make it into the model or were not significant in model evaluation and this was most likely due to these composite parameters not bringing original predictive information into the model and being highly correlated with other predictors in the list. As previously notes, there is a mistake in the sign of LR75 (all positive values) in the dataset, so it should be labeled the minimum of LR75.

Predictor	**Deviance with Y**		**Predictor**	**Deviance with Y**
SHERB35	675		**LR36**	720
SHERB3	676		DPT2M	722
U15SHR	680		**U1SHR**	722
U2SHR	681		**U2SHR**	723
MOSHE	683		TKE875	724
SHERB3	685		PWAT	725
SHERB35	689		U05SHR	725
MOSH	689		**LR35**	725
U25SHR	696		LR75	726
TKE925	697		TMPLFC	726
TKE950	697		U10M	726
UESHR	703		TMP500	727
SPFH1815	704		**DIVG650**	727
SHERBE	706		**SPFH775**	728
U1SHR	707		**U25SHR**	728
SHERBE35	707		**U3SHR**	729
HPBLSFC	708		**SPFH750**	729
U3SHR	708		**U10M**	730
PWAT	711		**SPFH1290**	730
HGTTRP	713		**SPFH800**	731
TKE900	713		**TMP500**	731
SPFH1512	715		**U05SHR**	731
MOSH	717		DIVG300	732
U15SHR	718		**S3MG**	733
CLBPRS	720		EFFSHR	733

Table 2.12. Top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR maximum and minimum value "severe wind speed" dataset (1198 cases). The value of each predictor was taken within a sample area (162x162 km^2). The **minimum value** predictors are bolded. Because this same dataset was used to develop the MOSH and MOSHE, it is no surprise that these two predictors made the top of this list. All predictors in the predictor list that are HSLC severe weather composite parameters (which all used the same or similar case list as this study) are among the most skillful predictors. However, all versions of the MOSH and SHERB predictors either did not make it into the model or were not significant in model evaluation to stay in the model, and this was most likely due to these composite parameters not bringing original predictive information into the model and being highly correlated with other predictors in the list. The same was true for all TEVV predictors, which is also a composite parameter. Of course, surface level wind (W10M, V10M, and U10M) is near the top of the skillful predictor list, because the event cases are reports of severe wind speeds (i.e., wind gust of 65 knots and faster); this is a good rationality check on the dataset.

Predictor	Deviance with Y		Predictor	Deviance with Y
TKE925	1401		SHERBE	1538
TKE950	1404		SHERBE35	1539
PMSL2	1410		TEVV5	1540
PMSL	1413		TEVV8	1541
MOSH	1439		HPBLSFC	1543
MOSHE	1447		TKE875	1544
PMSL2	1451		**DIVG875**	1548
PMSL	1454		TEVV45K	1549
SHERB3	1477		TEVV35K	1549
TKE900	1484		FGEN700	1550
SHERB35	1486		**TKE925**	1552
W10M	1490		TEVV6K	1552
V10M	1491		**W10M**	1553
TKE975	1509		TEVV4K	1555
TEVV2K	1514		FGEN725	1556
MAXTEVV	1517		FGEN750	1556
TEVV7	1518		**TEDF1085**	1556
TEVV25K	1519		FGEN775	1559
TKE1000	1523		**DIVG850**	1560
TKE950	1523		TEVV5K	1563
HPBLSFC	1524		**DIVG900**	1563
TEVV3K	1533		UESHR	1564
MINTEVV	1534		U10M	1565
SHERB35	1536		FGEN800	1567
SHERB3	1536		TKE850	1569

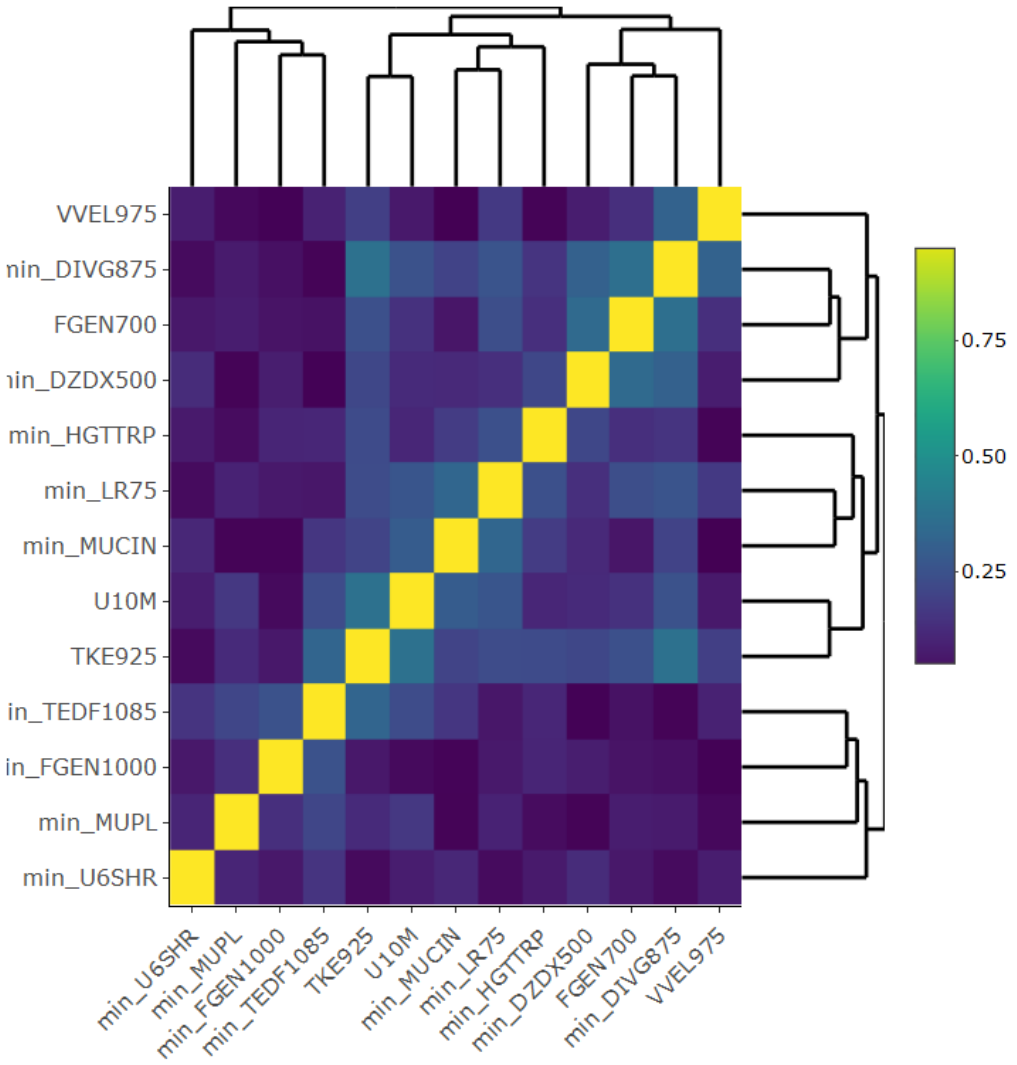


Figure 2.11. Correlation matrix and dendrogram for the 13 chosen predictors using the combined NARR maximum and minimum value "severe wind speed" dataset (1198 cases). The absolute value of predictor vs. predictor correlations is shown in the correlation matrix, where solid yellow represents a correlation of 1.00 (i.e., provides the same information) and solid blue represents a correlation of 0.00. Correlations of 0.4 and higher are considered highly correlated, so all of the predictors chosen have correlations less than 0.4 as shown. This gives an example of what the final correlation matrix looks like before the predictors are inputted into the logistic regression. Predictors with insignificant p-values are then removed from this list of predictors for the final model.

Table 2.13. Illustrative example of how a probabilistic model works as a holistic system. Because a predictor in the model is likely to increase/decrease at the same time of an increase/decrease of another predictor in the model, odds-ratios of a predictor may not give useful information. It is best to examine the probabilistic model as one complete system. The model used for this example was the probabilistic model produced using the combined dataset as the training dataset in section 2.6.4. Three cases were randomly chosen for an illustrative example; case 1 is a non-severe (unverified warning) environment (Y=0, predicted probability of 0.104), case 2 is a "more ambiguous" severe environment (Y=1, predicted probability of 0.600), and case 3 is a severe environment (Y=1, predicted probability of 0.983). The maximum/minimum predictor values taken from the NARR gridded data (32 km horizontal grid spacing) over a 5x5 grid boxes sample area is shown for each case. See Table 2.1 for more information on the predictors. As a reminder, UESHR, TKE925, and minimum TEDF1085 gave the largest contributions to this probabilistic model.

Predictor in Model (units)	**Case 1 Predictor Value**	**Case 2 Predictor Value**	**Case 3 Predictor Value**
min_HGTTRP (m)	16732	11031	12964
min_TEDF1085 (K/m)	-0.0211	-0.00148	0.00774
TKE925 (J/kg)	0.5000	2.8088	7.5527
min_FGEN1000 (K/[100 km*3h])	0.1490	-0.4264	-1.5806
FGEN725 (K/[100 km*3h])	0.9527	0.6855	3.8989
min_DIVG875 (s^{-1})	-0.0000287	-0.000109	-0.000123
UESHR (m/s)	17.28	14.22	28.35
NVWSTRP (s^{-1})	0.0086	0.0056	0.0063
U10M (m/s)	1.996	1.625	6.766
Predicted Probability	**0.104**	**0.600**	**0.983**

Table 2.14. The predictors and coefficients that make up the probabilistic model (from page 51) are shown.

Predictor (X_i)	**Coefficient (β_n)**
(intercept)	0.5809
HGTTRP	-0.0001363
U10M	0.09166
DIVG950	-3536
MOSHE	0.2894
FGEN800	0.1446
U2SHR	0.03666
DIVG250	5185
SBLI800	-0.1226

Table 2.15. The predictors and coefficients that make up the probabilistic model (from page 55) are shown.

Predictor (X_i)	**Coefficient (β_n)**
(intercept)	0.3826
HGTTRP	-0.0001353
MOSHE	0.2939
FGEN725	0.2054
SBLI800	-0.1076
U2SHR	0.03769
U10M	0.1003
DIVG250	4988

Table 2.16. The predictors and coefficients that make up the probabilistic model (from page 62) are shown.

Predictor (X_i)	**Coefficient (β_n)**
(intercept)	0.6389
TKE925	0.4387
UESHR	0.06880
FGEN725	0.2230
SBLI800	-0.06776
DIVG250	5059
HGTTRP	-0.0001659
MULFC	-0.0001135
DIVG950	-4769

Table 2.17. The predictors and coefficients that make up the probabilistic model (from page 67) are shown.

Predictor (X_i)	**Coefficient (β_n)**
(intercept)	-0.8601
min_HGTTRP	-0.0001493
min_TEDF1085	59.93
TKE925	0.2976
min_**FGEN1000**	0.1470
FGEN725	0.1814
min_**DIVG875**	-5115
UESHR	0.07692
NVWSTRP	60.32
U10M	0.06869

Table 2.18. The predictors and coefficients that make up the probabilistic model (from page 73) are shown.

Predictor (X_i)	**Coefficient (β_n)**
(intercept)	6.745
min_SPFH1815	-372.4
U1SHR	0.09213
min_DIVG650	-19270
min_TEDF1050	298.4
min_DIVG875	-11670
EFFSRH	0.002461
NSBCIN	-0.04658
V05SHR	0.08683
MUPL	-0.0001204
RH2M	-0.07427

Table 2.19. The predictors and coefficients that make up the probabilistic model (from page 77) are shown.

Predictor (X_i)	**Coefficient (β_n)**
(intercept)	-0.6037
min_TEDF1085	55.32
min_HGTTRP	-0.00007933
min_FGEN1000	0.2144
TKE925	0.3935
FGEN700	0.2106
min_DIVG875	-4782
min_DZDX500	-466000000
min_MUPL	-0.001441
min_U6SHR	-0.03275
U10M	0.05314

CHAPTER III

Creation of New Training Datasets for the Statistical Procedure

3.1. Introduction

This chapter discusses the statistical considerations and data collection steps taken to create appropriate training datasets for the severe weather statistical procedure (SWSP) that was outlined in section 2.5 regarding the prediction of HSLC severe weather hazards (research problem). These statistical considerations were proven necessary during the work conducted in Chapter II, and it is believed that similar data collection methods may be necessary when developing a meteorological dataset for training any severe weather statistical model. Chapter II results led to two hypotheses: 1) the probabilistic models will be improved by using smaller sample areas to calculate the maximum and minimum of predictor values in the training dataset, and 2) the probabilistic models will be improved by using higher resolution gridded model data (i.e., gridded model with smaller grid spacing and temporal spacing) for the predictor values in the training dataset. (See sections 2.6 and 2.7 for more details.) To test these two hypotheses, multiple new training datasets must be created. While creating these training datasets for the SWSP, statistical considerations should be taken into account (discussed in section 3.2).

The statistical considerations discussed in section 3.2 are rudimentary but have been often ignored during the development of previous severe weather prediction models, including severe weather composite parameters. These data collection methods were specifically designed for local scale (i.e., scale between regional scale and storm-scale) convective environments. Different or more complex methods may be necessary for storm-scale convective environments, because the training dataset would be developed from a convection-allowing numerical weather

prediction model. A convection-allowing model (CAM) is a weather forecast model that uses a horizontal grid spacing of 4 km or less, which explicitly predicts convective initiation and growth/decay without a convection parameterization scheme. A training dataset that is populated from a CAM that has a high enough temporal resolution (within minutes of the reported severe weather hazard) may not need maximum and minimum values of each variable (predictor); the actual predictor value calculated by the CAM in the nearest grid box to the report may be sufficient. (See section 6.4 for further discussion.)

The "master case list" from Chapter II was used to identify HSLC cases in the SEUS, and was trimmed down (i.e., specific cases were deleted) to meet the assumptions of the statistical techniques (that were used in SWSP) by minimizing correlations among the data values as discussed in section 3.2. For each case in the training datasets, the information given included: month, day, year, exact time of storm/warning report, latitude, longitude, and report/warning type. The NARR and RAP gridded model data, discussed in section 3.3, were determined to be the best options for representing the atmospheric environment of local severe storms for all cases. The training datasets were populated from these model datasets as discussed in section 3.4. To account for the storm motion between the time of the storm/null report (i.e., event/null case) and the time of the closest NARR/RAP model data point, a sample radius was calculated for each dataset to properly sample the local storm environment. Too large of a radius can lower the resolution of the predictor values in the dataset, and too high of a time difference between storm report and NARR/RAP model time stamp can lower the skill of the probabilistic model. Therefore, a smaller 3x3 dataset was also created in addition to the corresponding 5x5 dataset for comparisons. All datasets were produced utilizing NCAR supercomputers, Yellowstone and Cheyenne (CISL 2017, 2018). The produced datasets were put into CSV spreadsheets, an

understandable format for the R program (R Core Team 2017) to read. The data spreadsheets were then used as the training datasets (i.e., input data) for the statistical models, as discussed in Chapter IV.

3.2. Statistical Considerations for Data Collection

After collecting all the available cases (i.e., master case list) within the region of interest (e.g., SEUS) during specified environmental conditions (e.g., HSLC convective environments), it is important to filter out cases that could cause correlation errors in the statistical models. In the SWSP, logistic regression is used to produce a statistical model that predicts the probability of a tornado and/or severe wind speed, i.e., a 100% probability of a tornado and/or severe wind speed is labeled as Y=1 and a 0% probability of a tornado and/or severe wind speed is labeled as Y=0. As noted, the response variable (Y) in the statistical model is categorical, either 1 or 0. To determine which model is most skillful, precautionary steps need to be followed to avoid spatial correlation, temporal correlation, and multicollinearity.

Generally, when using meteorological data, if a surface station is too close to another surface station, these surface observations (i.e., predictor values) can have a spatial correlation, which will introduce errors into the model. The variables (i.e., predictors) that are sampled in close proximity will have similar patterns, which can lower p-values and possibly result in a significant p-value because the overlapping information falsely accounts for more explained variation in the response variable. If a spatial correlation is present, then the data from one of these surface stations should be removed from the dataset since those variabilities are already being accounted for in the model via the other surface station data. Specifically, when using gridded model data (e.g., RAP) to populate the datasets, spatial correlations can occur among the

sample areas (i.e., cases) that are substantially overlapping. One maximum value and one minimum value for each predictor (i.e., predictor values) are collected from each sample area, and either value could be the same value in both samples if the sample areas are overlapped. Therefore, for this research, each sample area is designed to not overlap another sample area, by disregarding close proximity cases as discussed in section 3.4. Note that in general, the cases (i.e., samples) cannot be considered completely independent because spatial data points cannot be considered independent from each other; therefore, we cannot assume independence of samples. However, creating spatial separation between samples (section 3.4) can help minimize the spatial correlation errors.

Temporal correlation can be induced when collecting surface data at regular intervals from the same surface station. Surface observations that are collected sequentially over time create a temporal correlation between data points because observations near each other in time will be similar; therefore, the assumption that observations are independent is violated in this scenario. When using time series data, temporal correlation can lead to predictors appearing to have less variance among each other, which affects the calculated p-values and estimated coefficients for each predictor kept in the model. This is because the redundant (overlapping) information produces similarities (patterns) between observations that do not naturally exist, artificially lowering the p-value and possibly creating a significant p-value. Using a sufficiently low sampling frequency at each surface station can avoid this problem; therefore, this research minimizes the number of samples (i.e., cases) from the same date and time as discussed in section 3.4.

As mentioned in Chapter II, multicollinearity in a regression model occurs when two or more predictors are highly correlated among each other. In this research, the response variable is

categorical, so multicollinearity can easily distort the interpretation of the model and can lead to unreliable coefficients in the logistic regression. A least squares model, such as a regression model, assumes there is no perfect collinearity, which means there is no exact linear relationship among the predictor variables. Essentially, multicollinearity causes difficulties in computing the least squares estimates by preventing the procedure from isolating and measuring the contribution of each predictor variable on the response variable. Small changes in the input data (i.e., predictor values) can lead to large changes in the final statistical model, which would affect the parameter estimates (coefficients) and may even change the sign of the coefficients. In the case of multicollinearity, the "duplicate" predictor(s) must be dropped from the statistical model. To avoid this issue in this research study, predictors that have a correlation of 0.4 or higher with another predictor are not included in the model (as discussed in Step 7); this is done to minimize multicollinearity and limit the number of chosen predictors. This correlation threshold was determined through dataset evaluation, model evaluation, and also recommended by Alfaro-Cordoba (2017).

Although including more cases in the training dataset can produce more precise coefficients in the model, using all the cases in the master case list introduces temporal and spatial correlations, as discussed in section 2.7. A predictor value that has no significant relationship with the response variable by itself could show a statistically significant p-value when combined with other highly correlated variables in the model, resulting in a false positive (i.e., Type I error). It is difficult to distinguish the individual contributions of each variable when multicollinearity exists due to the overlap between the contributions of the highly correlated variables. Alternatively, multicollinearity could increase uncertainty (widening confidence intervals) so the p-value would incorrectly show no statistical significance, resulting in a false

negative result (i.e., Type II error). The weight of the predictors (coefficients) will also be affected if any false correlations (i.e., spatial, temporal or multicollinearity) are present. These correlations among the dataset create patterns that are not physically present and are due to the false correlation, which can inflate the weight of those predictor variables. If there is any type of false correlation among the predictor variables, the model may incorrectly produce significant p-values (i.e., p-values that are too small) and keep predictors in the model that are not highly correlated (according to the chosen cut-off alpha) with the response variable. These issues may also potentially remove useful predictors from the model. If a predictor that accounts for some of the variability in the response (Y) is removed from the model, that variability is added to the error term (i.e., estimated error variance). This can cause a correlation between the error term and the predictors, which violates the assumption that all explanatory variables are uncorrelated with the error term. This also means that the estimated coefficients of the predictors left in the model will be too high, which is called omitted variables bias. Overall, any correlation error can result in a final model that is not skillful in predicting the phenomenon of interest (response variable), which in this research is the probability of a tornado and/or severe wind speed.

One last consideration, which is one of the most common statistical errors in research studies (Gigerenzer 2004, Ziliak and McCloskey 2009), is that a significant p-value does not equate to skill. For example, a probabilistic model consisting of all significant predictors (i.e., predictors with significant p-values) may not be a skillful model. The probabilistic model needs to go through multiple validation tests to determine true model skill (see section 2.5); this includes checking for model assumption violations, which can erroneously lead to a higher model skill when present. In any circumstance, p-values reported without context are meaningless. In addition, p-values of predictors cannot be compared between different training

datasets; for example, two training datasets with the same case list but different predictor values, such as values calculated with a different horizontal grid spacing. This means probabilistic models that were trained with different datasets cannot be compared using AIC or deviance.

By statistical definition, a significant p-value occurs when the null hypothesis is rejected. In this research, the null hypothesis was defined as there being no significant difference between the predictor values (e.g., surface CAPE values) in the null cases compared to the corresponding predictor values in the event cases. If there is a significant difference between these values (i.e., a significant p-value is calculated for the predictor), then the predictor is useful in discriminating between null and event cases, to some extent. If we were to use only one significant predictor in a model, it is unlikely the model would be skillful when validated. Especially with meteorological data, a specific combination of significant predictors is needed to produce a skillful model; how to determine the specific combination of predictors is discussed in section 2.5. This statement does not include composite parameters, even though they are technically an individual predictor in the model (i.e., gives one value), because a composite parameter is comprised of multiple skillful variables that essentially give the composite parameter skill from multiple individual predictors. However, as shown in section 2.6, it is better to include the significant variables as individual predictors with appropriate weights (coefficients) in a probabilistic model, rather than to create a probabilistic model with just one predictor comprised of all significant variables (i.e., composite parameter). This is one reason why severe weather composite parameters are not included in the predictor lists described in section 3.3.

3.3. Gridded Model Data

The intention of this research was to provide probabilistic models for operational forecasting, thus it was a priority to also use a numerical weather prediction (NWP) model that is currently used by severe weather researchers and forecasters. Therefore, data from two gridded models, the North American Regional Reanalysis (NARR) and the 13-km Rapid Refresh (RAP), were used in this study. The NARR model data is fundamentally a reanalysis dataset that was generated from a now-retired NWP model plus data assimilation system. More specifically, the NARR was described by the developers as an extension of the NCEP Global Reanalysis that covers North America and uses the NCEP Eta Model (32km/45 layer) together with the Regional Data Assimilation System (Mesinger et al. 2006, NARR 2016). The NARR model data has a horizontal grid spacing of approximately 32 km (per grid box) and temporal spacing of 3 hours. NARR data is output at 29 vertical pressure levels (i.e., 1000, 975, 950, 925, 900, 875, 850, 825, 800, 775, 750, 725, 700, 650, 600, 550, 500, 450, 400, 350, 300, 275, 250, 225, 200, 175, 150, 125, 100). The NARR case dates ranged from January 2006 to May 2014.

The RAP, currently an operational NWP model at NOAA NCEP, was described by the developers as an hourly-updated modeling system covering North America that comprises of a numerical forecast model and an analysis/assimilation system to initialize that model (RAP 2018). The RAP model replaced the Rapid Update Cycle (RUC) model in May 2012; therefore, all RAP data before May 2012 is technically RUC data, which used the same model structure, so it is commonly referred to as the RAP (Benjamin et al. 2004, 2016). Model output from both the RAP and RUC (Grid 130) have a horizontal grid spacing of approximately 13 km (per grid box) and temporal spacing of 1 hour. Only RUC data with a 13 km grid spacing was used, so the cases in the “RAP” training datasets with dates before April 2007 were deleted; the “RAP” case dates

ranged from April 1, 2007 to the end of 2015. There were also 259 other cases that were deleted due to missing 13 km RAP analysis data in the data archives. Due to the data collection and processing methodology, 20 km RAP analysis was not used as a substitute. RAP data is output at 37 vertical pressure levels (i.e., every 25 hPa from 1000 hPa to 100 hPa). Specifically, the analysis (i.e., 0-hour forecast) data was used from both the RAP and RUC in this study; a reanalysis dataset was not available.

Analysis data provides variable values that are the closest to observational values, compared to forecast data. However, the analysis system used in an NWP model can change (be updated by developers) over the years, so there is a possibility of varying model bias throughout the 2007 to 2015 "RAP" analysis training dataset. The NARR data are specifically reanalysis data, which provide predictor values from a uniform model (i.e., one version of the Eta model, thus no model updates) for all years in the case list. Using reanalysis data minimizes concern for (NWP) model bias error, and the NARR is currently the highest resolution reanalysis dataset available. The variables (predictors) in the NARR training datasets are shown in Table 3.1. To avoid extra computational time for the user of the probabilistic models (e.g., NWS forecaster), additional variables were not calculated from the NARR data (as it was done in Chapter II). This decision was also for the reason that calculating additional variables can add more highly correlated predictors to the dataset (i.e., multicollinearity), and the goal of this research was to find a combination of skillful predictors to include in a probabilistic model, not to design a new composite parameter (i.e., one variable made up of multiple variables). During the Chapter II investigation (section 2.6), it was found that when a composite parameter is used as a single predictor in a probabilistic model, it can produce a skillful model; however, a probabilistic model

comprised of the composite parameter's individual variables (separate predictors) can produce a more skillful model. Therefore, composite parameters were not calculated.

The RAP is currently used in severe weather forecasting operations at the NWS Storm Prediction Center and local NWS Weather Forecasting Offices. It is also used to initialize the High-Resolution Rapid Refresh (HRRR) model, which is another higher resolution (3 km, 15-minute) NWP model used for severe weather prediction. The RAP model was chosen because it is currently the highest resolution NWP model used by NWS operational forecasters that is available for the entire time period (of the master case list) and not a CAM, which would involve more data collection considerations. The list of common variables available in both the RUC and RAP that were used is shown in Table 3.2, and these variables are referred to as the RAP model predictors. Only the variables available in both the RUC and RAP were used for the variable list so all cases (years) had the same predictors throughout the "RAP" training dataset, i.e., a uniform dataset. For the same reasons given above for the NARR variable list, additional variables were not calculated from the RAP.

The Eta (NARR) and RAP models are considered regional (mesoscale) models, so both models can represent the "local" storm environment. Note that this is not the same as the "storm-scale" environment, which would require a much higher resolution; however, this research study was focused on studying the environment in the regional-to-local spatial scale and hour-to-minute temporal scale. There are numerous reasons for producing multiple statistical models in this study, including the ability to compare statistical models produced with different input data resolutions. NARR data have a lower resolution (i.e., 32 km and 3-hourly) compared to the RAP data (i.e., 13 km and 1-hourly), and these comparisons can provide insight into the optimal resolution of the training dataset for the statistical models. Comparisons between the NARR and

RAP statistical models can determine if it is better or worse to use a higher resolution dataset when predicting severe weather hazards (i.e., tornadoes and severe wind speeds), which was one of the scientific questions of this study.

3.4. Creation of Training Datasets for Statistical Procedure

The case list for all null and event cases in this study was taken from Sherburn et al. (2016), which is labeled the "master case list" in this dissertation research. As a reminder, all cases had storm environments that met the criteria of high-shear, low-CAPE environments, and all cases occurred in the Southeastern United States as described in section 2.3. For each training dataset, cases from the master case list were selected based on the criteria described in sections 3.4.1 and 3.4.2. The training datasets were then populated using numerical weather prediction output from the NARR or RAP as described in this chapter. The methods described in this section (3.4) were designed for statistical reasons, as well as to maximize the amount of local storm environment information included in training the statistical models while also minimizing including too much area outside of this local storm environment, which can muddle the predictive information. The statistical reasoning was necessary to ensure the datasets were appropriate training datasets for statistical modeling, which included minimizing spatial and temporal correlation errors in the statistical models. Section 3.2 has more information on correlation errors in statistical modeling. Details on the training datasets are shown in Table 3.3, including the name of the gridded model data (NARR or RAP), estimated number of grid boxes (3x3 or 5x5), sample radius, maximum time difference between case time and gridded model time, number of event cases, and number of null cases.

3.4.1. Calculations to Determine NARR Case Lists & Sample Radius

NARR model data was available temporally every 3 hours, so the closest model data point is up to 1.5 hours off the time of the case (event or null). The closest NARR model time for each case was found, and then the case list was sorted based on the time difference between the NARR model time and the case time (i.e., temporal difference). This temporal difference was used to filter the case list for each dataset (i.e., NARR 5x5 and NARR 3x3). NARR data is available spatially every 32.4 km (NOAA NCEP 2017), and the 32.4 km grid box length is equivalent to 0.292 degrees latitude or longitude, assuming 1 degree equals 111 km, which is a good estimate for the mid-latitudes (e.g., Southeastern United States). The 0.292 degrees value was used to determine the sample area, which consisted of either 5x5 or 3x3 grid boxes.

With the temporal difference in mind, a maximum storm motion was chosen for calculations to ensure the storm cell was still within the sample area at the NARR model time. Unfortunately, there are no known studies that have estimated the average or maximum storm motion speed of HSLC storm cells in the SEUS. Based on radar measurements, using the WSR-88D SCIT algorithm (Johnson et al. 1998), of real-time observations of HSLC storms within the SEUS over a one-year period by the author, a maximum storm motion of 35 knots was chosen (i.e., 35-knot storm motion assumption) because it represented around 95% of the observed storm cells. To justify these observations, this 35-knot value was compared with other research studies that calculated storm motion speeds. Edwards et al. (2002) calculated mean storm motion speeds around 20 knots for weakly tornadic (weaktor) and non-tornadic (nontor) supercells using real-time WSR-88D reflectivity displays, following the storm centroids through a series of volume scans lasting > 30 minutes per storm (Figure 3.1). About 1,862 out of 2,063 HSLC cases in the “master case list” fit into "weaktor" and "nontor" categories. Most of the HSLC cases are

also not supercellular, which would potentially lead to lower storm motion speeds for the HSLC cases than those reported by Edwards et al. (2002). The mean cloud top height and cloud depth of HSLC storms are comparably lower than the supercell storms previously studied (Davis and Parker 2014), which can also contribute to lower storm motion speeds than those calculated by Edwards et al. (2002). Therefore, if the mean storm motion speed of HSLC storms was closer to 15 knots, a max of 35 knots would be reasonable, considering a Gaussian distribution. It is noted that there would be outliers of faster storm motion speeds, but it would be a small percentage of cases.

For the NARR 5x5 dataset, the case list was first filtered based on the temporal difference described above to keep only the cases that occurred within 1 hour or less of the NARR model time. The length of five grid boxes is equivalent to 1.46 degrees; therefore, any case (i.e., sample) within 1.46 degrees or less of another case was deleted from the case list. This allowed for a case list with no overlap of samples, which minimized spatial correlation error in the statistical models. Cases were deleted if there were more than two cases with the same year, month, day, and hour, to minimize temporal correlation error in the statistical models. When deciding on which case to delete, the case with the largest temporal difference from the model time was deleted. This protocol to minimize spatial and temporal correlation errors was followed for all training datasets (i.e., NARR 5x5, NARR 3x3, RAP 5x5, and RAP 3x3).

Using the 35-knot storm motion assumption, storms travel up to 64.8 km (0.584 degrees) in one hour, so a sample radius of 64.8 km (for the NARR 5x5) was chosen (Figure 3.2). This radius was drawn around the point location (i.e., latitude and longitude) of the event/null case (i.e., sample) using the great circle distance function in MATLAB (MATLAB 2014), which calculates the shortest arc length between two points on a sphere (i.e., Earth). Because the storms

within the case list propagated in all cardinal directions, all data within a 64.8 km radius around the point location was included in the sample. Therefore, this radius would sample the storm environment within a sample area corresponding to 5x5 grid boxes. There is a benefit to using the sample radius instead of the actual 5x5 grid boxes, which is due to the grid boxes being stretched out asymmetrically over a Lambert Conformal Conic grid (used by the NARR and RAP). Due to the shape of the grid boxes, an individual grid box within the 5x5 sample area will not be included in the sample if the sample radius does not reach the center of that grid box (Figure 3.3). In other words, a 5x5 sample area could consist of less than 25 grid boxes, and a 3x3 sample area could consist of less than 9 grid boxes (e.g., Figure 3.4). This allows for a smaller sample region that still captures the local storm environment while minimizing data (influence) from outside the immediate local storm environment. Note that if it is found that a the 35-knot storm motion assumption is not fast enough, this storm motion speed assumption can be increased up to 52 knots (60 mph) and the corresponding increased sample radiuses will still fit within the 3x3 and 5x5 grid boxes areas. As long as the storm cells are being captured within the sample area, the smaller the sample radius, the more robust the training dataset. (See section 4.2.3 which tested this sensitivity.)

For the NARR 3x3 dataset, the case list was first filtered based on the temporal difference described above to keep only the cases that occurred within 30 minutes or less of the NARR model time. The length of three grid boxes is equivalent to 0.876 degrees; therefore, any case (i.e., sample) within 0.876 degrees or less of another case was deleted from the case list (for no overlap of samples). The protocol to minimize spatial and temporal correlation errors, as discussed with the NARR 5x5 dataset, was followed. Using the 35-knot storm motion assumption, storms travel up to 32.4 km (0.292 degrees) in 30 minutes, so a sample radius of

32.4 km was chosen (Figure 3.2). This sample radius was used in the same way as discussed with the NARR 5x5 dataset. Therefore, this radius would sample the storm environment within a sample area corresponding to 3x3 grid boxes.

3.4.2. Calculations to Determine RAP Case Lists & Sample Radius

RAP model data was available temporally every 1 hour, so the closest model data point is up to 30 minutes off the time of the case (event or null). The closest RAP model time for each case was found, and then the case list was sorted based on the time difference between the RAP model time and the case time (i.e., temporal difference). This temporal difference was used to filter the case list for each dataset (i.e., RAP 5x5 and RAP 3x3). RAP data is available spatially every 13.5 km (NOAA NCEP 2017), and the 13.5 km grid box length is equivalent to 0.122 degrees latitude or longitude, assuming 1 degree equals 111 km. The 0.122 degrees value was used to determine the sample area, which consisted of either 5x5 or 3x3 grid boxes.

For the RAP 5x5 dataset, the case list was first filtered based on the temporal difference described above to keep only the cases that occurred within 30 minutes or less of the RAP model time. The length of five grid boxes is equivalent to 0.61 degrees; therefore, any case (i.e., sample) within 0.61 degrees or less of another case was deleted from the case list (for no overlap of samples). The protocol to minimize spatial and temporal correlation errors, as discussed with the NARR 5x5 dataset, was followed. Using the 35-knot storm motion assumption described in section 3.4.1, storms travel up to 32.4 km (0.292 degrees) in 30 minutes, so a sample radius of 32.4 km was chosen (Figure 3.2). This sample radius was used in the same way as discussed with the NARR 5x5 dataset. Therefore, this radius would sample the storm environment within a sample area corresponding to 5x5 grid boxes.

For the RAP 3x3 dataset, the case list was first filtered based on the temporal difference described above to keep only the cases that occurred within 15 minutes or less of the RAP model time. The length of three grid boxes is equivalent to 0.366 degrees; therefore, any case (i.e., sample) within 0.366 degrees or less of another case was deleted from the case list (for no overlap of samples). The protocol to minimize spatial and temporal correlation errors, as discussed with the NARR 5x5 dataset, was followed. Using the 35-knot storm motion assumption, storms travel up to 16.2 km (0.146 degrees) in 15 minutes, so a sample radius of 16.2 km was chosen (Figure 3.2). This sample radius was used in the same way as discussed with the NARR 5x5 dataset. Therefore, this radius would sample the storm environment within a sample area corresponding to 3x3 grid boxes.

3.4.3. Populating Datasets with Predictor Values

As discussed, all new training datasets had sample areas of approximately 9 (3x3) or 25 (5x5) grid boxes (usually less) which allowed for multiple grid boxes to be sampled for each predictor value. This was to consider not only storm motion but also irregularities in the locations of storm development and propagation within the gridded (NWP) model, which may not well represent the observed atmospheric conditions. Because the case times were one hour or less from the gridded model time and the gridded model data used was (re)analysis data (which nudged the gridded model data closer to observations), it was believed that the storm environment related to the event or null was reasonably represented somewhere within the sample area.

To represent the storm environment within the sample area, the maximum and minimum values for each variable in the gridded model (i.e., NARR or RAP) were calculated. The

maximum and minimum statistics were chosen because they represent the extreme values of each predictor within the sample area (i.e., local storm environment for a specific case). Taking a different statistic, such as a mean value, was not advised due to the large size of the sample areas (up to 65 km radius). For example, if a large CAPE value is necessary for tornadogenesis, taking the average of all CAPE values within the sample area (i.e., the mean) would lower the CAPE predictor value; therefore, CAPE may not show up as a skillful predictor. As noted in sections 3.4.1 and 3.4.2, four datasets were computed (i.e., NARR 5x5, NARR 3x3, RAP 5x5, and RAP 3x3). Each dataset contained maximum values for each predictor (i.e., NARR/RAP variable) as well as minimum values for each predictor. The datasets were used as the input data, also known as the training datasets, for the SWSP discussed in section 2.5.

3.5. Summary

Statistical considerations were carefully thought-out for data collection methods to create training datasets for the SWSP, which involved calculating predictor values within a specified sample area of a gridded model. These considerations may not be necessary for data collection methods that involve one grid point, e.g., higher resolution storm-centered methods that follow the center of the storm. Storm-centered datasets are not available, and may not currently exist, for a long enough time period needed to train a probabilistic model. Minimizing the costs and resources of running the probabilistic models in NWS operations was also taken into consideration during the development of the input (training) datasets for the SWSP, which included not calculating additional predictors. The spatial and temporal definitions chosen for the training datasets were a consequence of focusing on the mesoscale-to-local scale severe environment and the typical storm motion of HSLC severe storms (in both magnitude and

direction). These training datasets focused on the time before a severe weather hazard of 15-minute, 30-minute, and 1-hour to try to have samples (cases) that represented the storm environment when the severe weather hazards occurred. If a warning is not verified (i.e., a severe weather hazard was not recorded), it is labeled as a false alarm, which was a major focus of this research. One way to minimize false alarms is by providing better severe weather forecasts in the 0-hour to 1-hour lead times. Tornado warnings and severe weather warnings are typically issued during these time periods. Once skillful probabilistic models are produced for the 0-hour (analysis), they can be tested for forecast hours further out in advance to test the sensitivity of the probabilistic model skill on increasing uncertainty of predictor values in the gridded (NWP) models (see section 5.3).

A RAP 1x1 (one grid box) dataset was created with a temporal difference of 5 minutes or less between the case time and the RAP model time, and a sample radius of 6 km. The RAP 1x1 dataset only contained 229 event cases and 120 null cases, which was not a large enough sample size to train a statistical model. The official storm report times, in general, were not expected to be precise enough for this temporal difference to be significantly different from the 15 minutes or less temporal difference of the RAP 3x3 dataset, so the RAP 1x1 dataset was not analyzed. It would also be difficult for the RAP model to recreate the observed storm environment in the exact location of the observed storm report (i.e., the same grid box) for all cases. Therefore, cases within less than 15 minutes of the gridded model time could not be examined in this study due to the limited data available (not enough cases occurred less than 15 minutes apart from the gridded model data), which is an example of the need for long-term higher resolution NWP models for severe weather research. The HRRR model outputs data every 15 minutes, so when

enough years (cases) are available, this NWP model could be used to examine cases within less than 15 minutes (see section 6.4).

If the probabilistic models produced by the NARR data (i.e., NARR models) are similar in skill to the probabilistic models produced by the RAP data (i.e., RAP models), both sets of models can be provided to the NWS for operational testing. The RAP models can be tested with real-time RAP forecast data, and the NARR models could be reasonably tested with the RAP forecast data and other NWP model forecast data. Other studies (e.g., Sherburn et al. 2016) have developed composite parameters or prediction equations using a NARR training dataset and then successfully used those equations with RAP and other NWP data in real-time forecasting operations. The NARR and RAP probabilistic models can be tested using a random sample of cases in the dataset (testing dataset) for validation, and then tested using the entire dataset for the confusion matrix calculations. The models from different datasets (e.g., NARR 5x5) can be compared using the confusion matrix (to calculate FAR and POD at the optimal cutoff of the model) and AUROC (overall model skill). These results are reported in Chapter IV.

The data collection methods described in this chapter not only produced more reliable training datasets, they also provided a more interpretable result from the probabilistic models by giving spatial and temporal definitions for the resulting prediction (probability). Reporting a prediction with a specified radius (km) is more understandable than an area defined by grid boxes, and this is also the convention in severe storm prediction. Note that the spatial and temporal definitions do not limit the possible skill of these probabilistic models if used further in advance (e.g., 2-hour forecast time instead of 1-hour forecast time); see section 5.3. It is possible that similar data collection methods may be necessary when developing a meteorological dataset

for the purpose of training any severe weather probabilistic model; however, testing this hypothesis is outside the scope of this dissertation.

Table 3.1. Predictor list in the NARR (reanalysis) training datasets; all 251 predictors are shown individually for transparency. Note that some variables available in the NARR were not used in the predictor list such as soil, albedo, radiation flux, precipitation, and 3-hour accumulation/average variables. Based on the values in the datasets, the CLWMR variables from 550-100 hPa and ICMR variables from 600-1000 hPa were removed. Based on the predictor evaluations, horizontal moisture divergence and the layer calculations for SPFH and VVEL were removed because these predictors were highly correlated with the raw predictors and/or were not skillful on their own. Units shown in brackets.

Predictor	Level(s)	Full Name of Predictor [units]
HPBL	from surface	Planetary boundary layer height [m]
HGT	100 hPa	Geopotential height [gpm]
TMP	100 hPa	Temperature [K]
SPFH	100 hPa	Specific humidity [kg/kg]
VVEL	100 hPa	Pressure vertical velocity [Pa/s]
UGRD	100 hPa	u-component of wind [m/s]
VGRD	100 hPa	v-component of wind [m/s]
ICMR	100 hPa	Ice mixing ratio [kg/kg]
HGT	125 hPa	Geopotential height [gpm]
TMP	125 hPa	Temperature [K]
SPFH	125 hPa	Specific humidity [kg/kg]
VVEL	125 hPa	Pressure vertical velocity [Pa/s]
UGRD	125 hPa	u-component of wind [m/s]
VGRD	125 hPa	v-component of wind [m/s]
ICMR	125 hPa	Ice mixing ratio [kg/kg]
HGT	150 hPa	Geopotential height [gpm]
TMP	150 hPa	Temperature [K]
SPFH	150 hPa	Specific humidity [kg/kg]
VVEL	150 hPa	Pressure vertical velocity [Pa/s]
UGRD	150 hPa	u-component of wind [m/s]
VGRD	150 hPa	v-component of wind [m/s]
ICMR	150 hPa	Ice mixing ratio [kg/kg]
HGT	175 hPa	Geopotential height [gpm]
TMP	175 hPa	Temperature [K]
SPFH	175 hPa	Specific humidity [kg/kg]
VVEL	175 hPa	Pressure vertical velocity [Pa/s]
UGRD	175 hPa	u-component of wind [m/s]
VGRD	175 hPa	v-component of wind [m/s]
ICMR	175 hPa	Ice mixing ratio [kg/kg]
HGT	200 hPa	Geopotential height [gpm]
TMP	200 hPa	Temperature [K]
SPFH	200 hPa	Specific humidity [kg/kg]
VVEL	200 hPa	Pressure vertical velocity [Pa/s]
UGRD	200 hPa	u-component of wind [m/s]
VGRD	200 hPa	v-component of wind [m/s]
ICMR	200 hPa	Ice mixing ratio [kg/kg]

Table 3.1. (Continued).

HGT	225 hPa	Geopotential height [gpm]
TMP	225 hPa	Temperature [K]
SPFH	225 hPa	Specific humidity [kg/kg]
VVEL	225 hPa	Pressure vertical velocity [Pa/s]
UGRD	225 hPa	u-component of wind [m/s]
VGRD	225 hPa	v-component of wind [m/s]
ICMR	225 hPa	Ice mixing ratio [kg/kg]
HGT	250 hPa	Geopotential height [gpm]
TMP	250 hPa	Temperature [K]
SPFH	250 hPa	Specific humidity [kg/kg]
VVEL	250 hPa	Pressure vertical velocity [Pa/s]
UGRD	250 hPa	u-component of wind [m/s]
VGRD	250 hPa	v-component of wind [m/s]
ICMR	250 hPa	Ice mixing ratio [kg/kg]
HGT	275 hPa	Geopotential height [gpm]
TMP	275 hPa	Temperature [K]
SPFH	275 hPa	Specific humidity [kg/kg]
VVEL	275 hPa	Pressure vertical velocity [Pa/s]
UGRD	275 hPa	u-component of wind [m/s]
VGRD	275 hPa	v-component of wind [m/s]
ICMR	275 hPa	Ice mixing ratio [kg/kg]
HGT	300 hPa	Geopotential height [gpm]
TMP	300 hPa	Temperature [K]
SPFH	300 hPa	Specific humidity [kg/kg]
VVEL	300 hPa	Pressure vertical velocity [Pa/s]
UGRD	300 hPa	u-component of wind [m/s]
VGRD	300 hPa	v-component of wind [m/s]
ICMR	300 hPa	Ice mixing ratio [kg/kg]
HGT	350 hPa	Geopotential height [gpm]
TMP	350 hPa	Temperature [K]
SPFH	350 hPa	Specific humidity [kg/kg]
VVEL	350 hPa	Pressure vertical velocity [Pa/s]
UGRD	350 hPa	u-component of wind [m/s]
VGRD	350 hPa	v-component of wind [m/s]
ICMR	350 hPa	Ice mixing ratio [kg/kg]
HGT	400 hPa	Geopotential height [gpm]
TMP	400 hPa	Temperature [K]
SPFH	400 hPa	Specific humidity [kg/kg]
VVEL	400 hPa	Pressure vertical velocity [Pa/s]
UGRD	400 hPa	u-component of wind [m/s]
VGRD	400 hPa	v-component of wind [m/s]
ICMR	400 hPa	Ice mixing ratio [kg/kg]

Table 3.1. (Continued).

HGT	450 hPa	Geopotential height [gpm]
TMP	450 hPa	Temperature [K]
SPFH	450 hPa	Specific humidity [kg/kg]
VVEL	450 hPa	Pressure vertical velocity [Pa/s]
UGRD	450 hPa	u-component of wind [m/s]
VGRD	450 hPa	v-component of wind [m/s]
ICMR	450 hPa	Ice mixing ratio [kg/kg]
HGT	500 hPa	Geopotential height [gpm]
TMP	500 hPa	Temperature [K]
SPFH	500 hPa	Specific humidity [kg/kg]
VVEL	500 hPa	Pressure vertical velocity [Pa/s]
UGRD	500 hPa	u-component of wind [m/s]
VGRD	500 hPa	v-component of wind [m/s]
ICMR	500 hPa	Ice mixing ratio [kg/kg]
HGT	550 hPa	Geopotential height [gpm]
TMP	550 hPa	Temperature [K]
SPFH	550 hPa	Specific humidity [kg/kg]
VVEL	550 hPa	Pressure vertical velocity [Pa/s]
UGRD	550 hPa	u-component of wind [m/s]
VGRD	550 hPa	v-component of wind [m/s]
ICMR	550 hPa	Ice mixing ratio [kg/kg]
HGT	600 hPa	Geopotential height [gpm]
TMP	600 hPa	Temperature [K]
SPFH	600 hPa	Specific humidity [kg/kg]
VVEL	600 hPa	Pressure vertical velocity [Pa/s]
UGRD	600 hPa	u-component of wind [m/s]
VGRD	600 hPa	v-component of wind [m/s]
TKE	600 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	600 hPa	Cloud water mixing ratio [kg/kg]
HGT	650 hPa	Geopotential height [gpm]
TMP	650 hPa	Temperature [K]
SPFH	650 hPa	Specific humidity [kg/kg]
VVEL	650 hPa	Pressure vertical velocity [Pa/s]
UGRD	650 hPa	u-component of wind [m/s]
VGRD	650 hPa	v-component of wind [m/s]
TKE	650 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	650 hPa	Cloud water mixing ratio [kg/kg]
HGT	700 hPa	Geopotential height [gpm]
TMP	700 hPa	Temperature [K]
SPFH	700 hPa	Specific humidity [kg/kg]
VVEL	700 hPa	Pressure vertical velocity [Pa/s]

Table 3.1. (Continued).

UGRD	700 hPa	u-component of wind [m/s]
VGRD	700 hPa	v-component of wind [m/s]
TKE	700 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	700 hPa	Cloud water mixing ratio [kg/kg]
HGT	725 hPa	Geopotential height [gpm]
TMP	725 hPa	Temperature [K]
SPFH	725 hPa	Specific humidity [kg/kg]
VVEL	725 hPa	Pressure vertical velocity [Pa/s]
UGRD	725 hPa	u-component of wind [m/s]
VGRD	725 hPa	v-component of wind [m/s]
TKE	725 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	725 hPa	Cloud water mixing ratio [kg/kg]
HGT	750 hPa	Geopotential height [gpm]
TMP	750 hPa	Temperature [K]
SPFH	750 hPa	Specific humidity [kg/kg]
VVEL	750 hPa	Pressure vertical velocity [Pa/s]
UGRD	750 hPa	u-component of wind [m/s]
VGRD	750 hPa	v-component of wind [m/s]
TKE	750 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	750 hPa	Cloud water mixing ratio [kg/kg]
HGT	775 hPa	Geopotential height [gpm]
TMP	775 hPa	Temperature [K]
SPFH	775 hPa	Specific humidity [kg/kg]
VVEL	775 hPa	Pressure vertical velocity [Pa/s]
UGRD	775 hPa	u-component of wind [m/s]
VGRD	775 hPa	v-component of wind [m/s]
TKE	775 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	775 hPa	Cloud water mixing ratio [kg/kg]
HGT	800 hPa	Geopotential height [gpm]
TMP	800 hPa	Temperature [K]
SPFH	800 hPa	Specific humidity [kg/kg]
VVEL	800 hPa	Pressure vertical velocity [Pa/s]
UGRD	800 hPa	u-component of wind [m/s]
VGRD	800 hPa	v-component of wind [m/s]
TKE	800 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	800 hPa	Cloud water mixing ratio [kg/kg]
HGT	825 hPa	Geopotential height [gpm]
TMP	825 hPa	Temperature [K]
SPFH	825 hPa	Specific humidity [kg/kg]
VVEL	825 hPa	Pressure vertical velocity [Pa/s]
UGRD	825 hPa	u-component of wind [m/s]
VGRD	825 hPa	v-component of wind [m/s]
TKE	825 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	825 hPa	Cloud water mixing ratio [kg/kg]
HGT	850 hPa	Geopotential height [gpm]

Table 3.1. (Continued).

TMP	850 hPa	Temperature [K]
SPFH	850 hPa	Specific humidity [kg/kg]
VVEL	850 hPa	Pressure vertical velocity [Pa/s]
UGRD	850 hPa	u-component of wind [m/s]
VGRD	850 hPa	v-component of wind [m/s]
TKE	850 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	850 hPa	Cloud water mixing ratio [kg/kg]
HGT	875 hPa	Geopotential height [gpm]
TMP	875 hPa	Temperature [K]
SPFH	875 hPa	Specific humidity [kg/kg]
VVEL	875 hPa	Pressure vertical velocity [Pa/s]
UGRD	875 hPa	u-component of wind [m/s]
VGRD	875 hPa	v-component of wind [m/s]
TKE	875 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	875 hPa	Cloud water mixing ratio [kg/kg]
HGT	900 hPa	Geopotential height [gpm]
TMP	900 hPa	Temperature [K]
SPFH	900 hPa	Specific humidity [kg/kg]
VVEL	900 hPa	Pressure vertical velocity [Pa/s]
UGRD	900 hPa	u-component of wind [m/s]
VGRD	900 hPa	v-component of wind [m/s]
TKE	900 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	900 hPa	Cloud water mixing ratio [kg/kg]
HGT	925 hPa	Geopotential height [gpm]
TMP	925 hPa	Temperature [K]
SPFH	925 hPa	Specific humidity [kg/kg]
VVEL	925 hPa	Pressure vertical velocity [Pa/s]
UGRD	925 hPa	u-component of wind [m/s]
VGRD	925 hPa	v-component of wind [m/s]
TKE	925 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	925 hPa	Cloud water mixing ratio [kg/kg]
HGT	950 hPa	Geopotential height [gpm]
TMP	950 hPa	Temperature [K]
SPFH	950 hPa	Specific humidity [kg/kg]
VVEL	950 hPa	Pressure vertical velocity [Pa/s]
UGRD	950 hPa	u-component of wind [m/s]
VGRD	950 hPa	v-component of wind [m/s]
TKE	950 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	950 hPa	Cloud water mixing ratio [kg/kg]
HGT	975 hPa	Geopotential height [gpm]
TMP	975 hPa	Temperature [K]
SPFH	975 hPa	Specific humidity [kg/kg]
VVEL	975 hPa	Pressure vertical velocity [Pa/s]
UGRD	975 hPa	u-component of wind [m/s]
VGRD	975 hPa	v-component of wind [m/s]
TKE	975 hPa	Turbulent Kinetic Energy [J/kg]

Table 3.1. (Continued).

CLWMR	975 hPa	Cloud water mixing ratio [kg/kg]
HGT	1000 hPa	Geopotential height [gpm]
TMP	1000 hPa	Temperature [K]
SPFH	1000 hPa	Specific humidity [kg/kg]
VVEL	1000 hPa	Pressure vertical velocity [Pa/s]
UGRD	1000 hPa	u-component of wind [m/s]
VGRD	1000 hPa	v-component of wind [m/s]
TKE	1000 hPa	Turbulent Kinetic Energy [J/kg]
CLWMR	1000 hPa	Cloud water mixing ratio [kg/kg]
PRES	2 m	Pressure [Pa]
TMP	2 m	Temperature [K]
POT	2 m	Potential Temperature [K]
DPT	2 m	Dew point Temperature [K]
RH	2 m	Relative humidity [%]
LFTX	500-1000 hPa	Surface lifted index [K]
CAPE	surface-based	SB Convective available potential energy [J/kg]
CIN	surface-based	SB Convective inhibition [J/kg]
CAPE	0-180 hPa AGL	ML Convective available potential energy [J/kg]
CIN	0-180 hPa AGL	ML Convective inhibition [J/kg]
PWAT	entire column	Precipitable water [kg/m^2]
PRES	cloud base	Pressure [Pa]
HGT	cloud base	Geopotential height [gpm]
PRES	cloud top	Pressure [Pa]
HGT	cloud top	Geopotential height [gpm]
TMP	cloud top	Temperature [K]
HLCY	3000-0 m	Storm relative helicity [m^2/s^2]
USTM	6000-0 m	u-component of storm motion [m/s]
VSTM	6000-0 m	v-component of storm motion [m/s]
PRES	tropopause	Pressure [Pa]
HGT	tropopause	Geopotential height [gpm]
TMP	tropopause	Temperature [K]
UGRD	tropopause	u-component of wind [m/s]
VGRD	tropopause	v-component of wind [m/s]
VWSH	tropopause	Vertical wind speed shear [1/s]
PRES	max wind level	Pressure [Pa]
HGT	max wind level	Geopotential height [gpm]
UGRD	max wind level	u-component of wind [m/s]
VGRD	max wind level	v-component of wind [m/s]
HGT	0°C isotherm	Geopotential height [gpm]
RH	0°C isotherm	Relative humidity [%]
PRES	LCL level	Pressure [Pa]

Table 3.2. Predictor list in the RAP analysis (RAP/RUC combined) training datasets; all 283 predictors shown for transparency. Only variables available in all RAP and RUC files were used, which was limiting. Units shown in brackets. L1 is surface level, L6 is maximum wind level, L7 is tropopause level, L4 and L204 are freezing level, L243 is convective cloud top level, and L247 is the equilibrium level.

Predictor	Level(s)	Full Name of Predictor [units]
HGT	1000 hPa	Geopotential height [gpm]
HGT	975 hPa	Geopotential height [gpm]
HGT	950 hPa	Geopotential height [gpm]
HGT	925 hPa	Geopotential height [gpm]
HGT	900 hPa	Geopotential height [gpm]
HGT	875 hPa	Geopotential height [gpm]
HGT	850 hPa	Geopotential height [gpm]
HGT	825 hPa	Geopotential height [gpm]
HGT	800 hPa	Geopotential height [gpm]
HGT	775 hPa	Geopotential height [gpm]
HGT	750 hPa	Geopotential height [gpm]
HGT	725 hPa	Geopotential height [gpm]
HGT	700 hPa	Geopotential height [gpm]
HGT	675 hPa	Geopotential height [gpm]
HGT	650 hPa	Geopotential height [gpm]
HGT	625 hPa	Geopotential height [gpm]
HGT	600 hPa	Geopotential height [gpm]
HGT	575 hPa	Geopotential height [gpm]
HGT	550 hPa	Geopotential height [gpm]
HGT	525 hPa	Geopotential height [gpm]
HGT	500 hPa	Geopotential height [gpm]
HGT	475 hPa	Geopotential height [gpm]
HGT	450 hPa	Geopotential height [gpm]
HGT	425 hPa	Geopotential height [gpm]
HGT	400 hPa	Geopotential height [gpm]
HGT	375 hPa	Geopotential height [gpm]
HGT	350 hPa	Geopotential height [gpm]
HGT	325 hPa	Geopotential height [gpm]
HGT	300 hPa	Geopotential height [gpm]
HGT	275 hPa	Geopotential height [gpm]
HGT	250 hPa	Geopotential height [gpm]
HGT	225 hPa	Geopotential height [gpm]
HGT	200 hPa	Geopotential height [gpm]
HGT	175 hPa	Geopotential height [gpm]
HGT	150 hPa	Geopotential height [gpm]
HGT	125 hPa	Geopotential height [gpm]
HGT	100 hPa	Geopotential height [gpm]
TMP	1000 hPa	Temperature [K]
TMP	975 hPa	Temperature [K]
TMP	950 hPa	Temperature [K]
TMP	925 hPa	Temperature [K]

Table 3.2. (Continued).

TMP	900 hPa	Temperature [K]
TMP	875 hPa	Temperature [K]
TMP	850 hPa	Temperature [K]
TMP	825 hPa	Temperature [K]
TMP	800 hPa	Temperature [K]
TMP	775 hPa	Temperature [K]
TMP	750 hPa	Temperature [K]
TMP	725 hPa	Temperature [K]
TMP	700 hPa	Temperature [K]
TMP	675 hPa	Temperature [K]
TMP	650 hPa	Temperature [K]
TMP	625 hPa	Temperature [K]
TMP	600 hPa	Temperature [K]
TMP	575 hPa	Temperature [K]
TMP	550 hPa	Temperature [K]
TMP	525 hPa	Temperature [K]
TMP	500 hPa	Temperature [K]
TMP	475 hPa	Temperature [K]
TMP	450 hPa	Temperature [K]
TMP	425 hPa	Temperature [K]
TMP	400 hPa	Temperature [K]
TMP	375 hPa	Temperature [K]
TMP	350 hPa	Temperature [K]
TMP	325 hPa	Temperature [K]
TMP	300 hPa	Temperature [K]
TMP	275 hPa	Temperature [K]
TMP	250 hPa	Temperature [K]
TMP	225 hPa	Temperature [K]
TMP	200 hPa	Temperature [K]
TMP	175 hPa	Temperature [K]
TMP	150 hPa	Temperature [K]
TMP	125 hPa	Temperature [K]
TMP	100 hPa	Temperature [K]
RH	1000 hPa	Relative humidity [%]
RH	975 hPa	Relative humidity [%]
RH	950 hPa	Relative humidity [%]
RH	925 hPa	Relative humidity [%]
RH	900 hPa	Relative humidity [%]
RH	875 hPa	Relative humidity [%]
RH	850 hPa	Relative humidity [%]
RH	825 hPa	Relative humidity [%]
RH	800 hPa	Relative humidity [%]
RH	775 hPa	Relative humidity [%]
RH	750 hPa	Relative humidity [%]
RH	725 hPa	Relative humidity [%]
RH	700 hPa	Relative humidity [%]
RH	675 hPa	Relative humidity [%]

Table 3.2. (Continued).

RH	650 hPa	Relative humidity [%]
RH	625 hPa	Relative humidity [%]
RH	600 hPa	Relative humidity [%]
RH	575 hPa	Relative humidity [%]
RH	550 hPa	Relative humidity [%]
RH	525 hPa	Relative humidity [%]
RH	500 hPa	Relative humidity [%]
RH	475 hPa	Relative humidity [%]
RH	450 hPa	Relative humidity [%]
RH	425 hPa	Relative humidity [%]
RH	400 hPa	Relative humidity [%]
RH	375 hPa	Relative humidity [%]
RH	350 hPa	Relative humidity [%]
RH	325 hPa	Relative humidity [%]
RH	300 hPa	Relative humidity [%]
RH	275 hPa	Relative humidity [%]
RH	250 hPa	Relative humidity [%]
RH	225 hPa	Relative humidity [%]
RH	200 hPa	Relative humidity [%]
RH	175 hPa	Relative humidity [%]
RH	150 hPa	Relative humidity [%]
RH	125 hPa	Relative humidity [%]
RH	100 hPa	Relative humidity [%]
UGRD	1000 hPa	u-component of wind [m/s]
UGRD	975 hPa	u-component of wind [m/s]
UGRD	950 hPa	u-component of wind [m/s]
UGRD	925 hPa	u-component of wind [m/s]
UGRD	900 hPa	u-component of wind [m/s]
UGRD	875 hPa	u-component of wind [m/s]
UGRD	850 hPa	u-component of wind [m/s]
UGRD	825 hPa	u-component of wind [m/s]
UGRD	800 hPa	u-component of wind [m/s]
UGRD	775 hPa	u-component of wind [m/s]
UGRD	750 hPa	u-component of wind [m/s]
UGRD	725 hPa	u-component of wind [m/s]
UGRD	700 hPa	u-component of wind [m/s]
UGRD	675 hPa	u-component of wind [m/s]
UGRD	650 hPa	u-component of wind [m/s]
UGRD	625 hPa	u-component of wind [m/s]
UGRD	600 hPa	u-component of wind [m/s]
UGRD	575 hPa	u-component of wind [m/s]
UGRD	550 hPa	u-component of wind [m/s]
UGRD	525 hPa	u-component of wind [m/s]
UGRD	500 hPa	u-component of wind [m/s]
UGRD	475 hPa	u-component of wind [m/s]
UGRD	450 hPa	u-component of wind [m/s]
UGRD	425 hPa	u-component of wind [m/s]

Table 3.2. (Continued).

UGRD	400 hPa	u-component of wind [m/s]
UGRD	375 hPa	u-component of wind [m/s]
UGRD	350 hPa	u-component of wind [m/s]
UGRD	325 hPa	u-component of wind [m/s]
UGRD	300 hPa	u-component of wind [m/s]
UGRD	275 hPa	u-component of wind [m/s]
UGRD	250 hPa	u-component of wind [m/s]
UGRD	225 hPa	u-component of wind [m/s]
UGRD	200 hPa	u-component of wind [m/s]
UGRD	175 hPa	u-component of wind [m/s]
UGRD	150 hPa	u-component of wind [m/s]
UGRD	125 hPa	u-component of wind [m/s]
UGRD	100 hPa	u-component of wind [m/s]
VGRD	1000 hPa	v-component of wind [m/s]
VGRD	975 hPa	v-component of wind [m/s]
VGRD	950 hPa	v-component of wind [m/s]
VGRD	925 hPa	v-component of wind [m/s]
VGRD	900 hPa	v-component of wind [m/s]
VGRD	875 hPa	v-component of wind [m/s]
VGRD	850 hPa	v-component of wind [m/s]
VGRD	825 hPa	v-component of wind [m/s]
VGRD	800 hPa	v-component of wind [m/s]
VGRD	775 hPa	v-component of wind [m/s]
VGRD	750 hPa	v-component of wind [m/s]
VGRD	725 hPa	v-component of wind [m/s]
VGRD	700 hPa	v-component of wind [m/s]
VGRD	675 hPa	v-component of wind [m/s]
VGRD	650 hPa	v-component of wind [m/s]
VGRD	625 hPa	v-component of wind [m/s]
VGRD	600 hPa	v-component of wind [m/s]
VGRD	575 hPa	v-component of wind [m/s]
VGRD	550 hPa	v-component of wind [m/s]
VGRD	525 hPa	v-component of wind [m/s]
VGRD	500 hPa	v-component of wind [m/s]
VGRD	475 hPa	v-component of wind [m/s]
VGRD	450 hPa	v-component of wind [m/s]
VGRD	425 hPa	v-component of wind [m/s]
VGRD	400 hPa	v-component of wind [m/s]
VGRD	375 hPa	v-component of wind [m/s]
VGRD	350 hPa	v-component of wind [m/s]
VGRD	325 hPa	v-component of wind [m/s]
VGRD	300 hPa	v-component of wind [m/s]
VGRD	275 hPa	v-component of wind [m/s]
VGRD	250 hPa	v-component of wind [m/s]
VGRD	225 hPa	v-component of wind [m/s]
VGRD	200 hPa	v-component of wind [m/s]
VGRD	175 hPa	v-component of wind [m/s]

Table 3.2. (Continued).

VGRD	150 hPa	v-component of wind [m/s]
VGRD	125 hPa	v-component of wind [m/s]
VGRD	100 hPa	v-component of wind [m/s]
VVEL	1000 hPa	Pressure vertical velocity [Pa/s]
VVEL	975 hPa	Pressure vertical velocity [Pa/s]
VVEL	950 hPa	Pressure vertical velocity [Pa/s]
VVEL	925 hPa	Pressure vertical velocity [Pa/s]
VVEL	900 hPa	Pressure vertical velocity [Pa/s]
VVEL	875 hPa	Pressure vertical velocity [Pa/s]
VVEL	850 hPa	Pressure vertical velocity [Pa/s]
VVEL	825 hPa	Pressure vertical velocity [Pa/s]
VVEL	800 hPa	Pressure vertical velocity [Pa/s]
VVEL	775 hPa	Pressure vertical velocity [Pa/s]
VVEL	750 hPa	Pressure vertical velocity [Pa/s]
VVEL	725 hPa	Pressure vertical velocity [Pa/s]
VVEL	700 hPa	Pressure vertical velocity [Pa/s]
VVEL	675 hPa	Pressure vertical velocity [Pa/s]
VVEL	650 hPa	Pressure vertical velocity [Pa/s]
VVEL	625 hPa	Pressure vertical velocity [Pa/s]
VVEL	600 hPa	Pressure vertical velocity [Pa/s]
VVEL	575 hPa	Pressure vertical velocity [Pa/s]
VVEL	550 hPa	Pressure vertical velocity [Pa/s]
VVEL	525 hPa	Pressure vertical velocity [Pa/s]
VVEL	500 hPa	Pressure vertical velocity [Pa/s]
VVEL	475 hPa	Pressure vertical velocity [Pa/s]
VVEL	450 hPa	Pressure vertical velocity [Pa/s]
VVEL	425 hPa	Pressure vertical velocity [Pa/s]
VVEL	400 hPa	Pressure vertical velocity [Pa/s]
VVEL	375 hPa	Pressure vertical velocity [Pa/s]
VVEL	350 hPa	Pressure vertical velocity [Pa/s]
VVEL	325 hPa	Pressure vertical velocity [Pa/s]
VVEL	300 hPa	Pressure vertical velocity [Pa/s]
VVEL	275 hPa	Pressure vertical velocity [Pa/s]
VVEL	250 hPa	Pressure vertical velocity [Pa/s]
VVEL	225 hPa	Pressure vertical velocity [Pa/s]
VVEL	200 hPa	Pressure vertical velocity [Pa/s]
VVEL	175 hPa	Pressure vertical velocity [Pa/s]
VVEL	150 hPa	Pressure vertical velocity [Pa/s]
VVEL	125 hPa	Pressure vertical velocity [Pa/s]
VVEL	100 hPa	Pressure vertical velocity [Pa/s]
PRES	Surface	Pressure [Pa]
PTEND	Surface	Pressure tendency [Pa/s]
POT	2 m AGL	Potential Temperature [K]
DPT	2 m AGL	Dewpoint Temperature [K]
DEPR	2 m AGL	Dewpoint depression [K]
TMP	2 m AGL	Temperature [K]

Table 3.2. (Continued).

UGRD	10 m AGL	u-component of wind [m/s]
VGRD	10 m AGL	v-component of wind [m/s]
RH	2 m AGL	Relative humidity [%]
SPFH	2 m AGL	Specific humidity [kg/kg]
EPOT	Surface	Pseudo-adiabatic pot. Temperature [K]
CAPE	Surface	Convective Avail. Pot. Energy [J/kg]
CIN	Surface	Convective inhibition [J/kg]
LFTX	Surface	Surface lifted index [K]
HLCY	Surface	Storm relative helicity [m^2/s^2]
PRES	0°C isotherm	Pressure [Pa]
HGT	0°C isotherm	Geopotential height [gpm]
RH	0°C isotherm	Relative humidity [%]
PRES	Tropopause	Pressure [Pa]
POT	Tropopause	Potential Temperature [K]
UGRD	Tropopause	u-component of wind [m/s]
VGRD	Tropopause	v-component of wind [m/s]
PRES	max wind level	Pressure [Pa]
UGRD	max wind level	u-component of wind [m/s]
VGRD	max wind level	v-component of wind [m/s]
TMP	30-0 hPa AGL	Temperature [K]
RH	30-0 hPa AGL	Relative humidity [%]
UGRD	30-0 hPa AGL	u-component of wind [m/s]
VGRD	30-0 hPa AGL	v-component of wind [m/s]
VVEL	30-0 hPa AGL	Pressure vertical velocity [Pa/s]
TMP	60-30 hPa AGL	Temperature [K]
RH	60-30 hPa AGL	Relative humidity [%]
UGRD	60-30 hPa AGL	u-component of wind [m/s]
VGRD	60-30 hPa AGL	v-component of wind [m/s]
VVEL	60-30 hPa AGL	Pressure vertical velocity [Pa/s]
TMP	90-60 hPa AGL	Temperature [K]
RH	90-60 hPa AGL	Relative humidity [%]
UGRD	90-60 hPa AGL	u-component of wind [m/s]
VGRD	90-60 hPa AGL	v-component of wind [m/s]
VVEL	90-60 hPa AGL	Pressure vertical velocity [Pa/s]
TMP	120-90 hPa AGL	Temperature [K]
RH	120-90 hPa AGL	Relative humidity [%]
UGRD	120-90 hPa AGL	u-component of wind [m/s]
VGRD	120-90 hPa AGL	v-component of wind [m/s]
VVEL	120-90 hPa AGL	Pressure vertical velocity [Pa/s]
TMP	150-120 hPa AGL	Temperature [K]
RH	150-120 hPa AGL	Relative humidity [%]
UGRD	150-120 hPa AGL	u-component of wind [m/s]
VGRD	150-120 hPa AGL	v-component of wind [m/s]
VVEL	150-120 hPa AGL	Pressure vertical velocity [Pa/s]
TMP	180-150 hPa AGL	Temperature [K]
RH	180-150 hPa AGL	Relative humidity [%]

Table 3.2. (Continued).

UGRD	180-150 hPa AGL	u-component of wind [m/s]
VGRD	180-150 hPa AGL	v-component of wind [m/s]
VVEL	180-150 hPa AGL	Pressure vertical velocity [Pa/s]
PRES	highest tropo freezing level	Pressure [Pa]
HGT	highest tropo freezing level	Geopotential height [gpm]
RH	highest tropo freezing level	Relative humidity [%]
PWAT	entire atmospheric column	Precipitable water [kg/m^2]
HPBL	Surface	Planetary boundary layer height [m]
PRES	max e-pot-temp level	Pressure [Pa]
HGT	convective cloud top	Geopotential height [gpm]
HGT	equilibrium level	Geopotential height [gpm]
TMP	Tropopause	Temperature [K]
ABSV	500 hPa	Absolute vorticity [/s]

Table 3.3. Details on the four training datasets are shown. Table shows each dataset with the name of gridded model data (NARR or RAP), estimated number of grid boxes (3x3 or 5x5), sample radius, maximum time difference between case time and gridded model time, number of event cases, number of null cases, number of total cases (events and nulls), and the number of cases randomly selected for the testing dataset. Because of smaller temporal spacing in the RAP (1-hourly) compared to the NARR (3-hourly), the RAP datasets have more cases. Also, cases were deleted within a dataset if there were more than two cases with the same year, month, day, and hour, to minimize temporal correlation error in the statistical models.

	NARR 5x5	**NARR 3x3**	**RAP 5x5**	**RAP 3x3**
sample radius	64.8 km	32.4 km	32.4 km	16.2 km
max time difference	1 hour	30 min	30 min	15 min
# of event cases	409	256	579	409
# of null cases	431	253	529	348
# of total cases	840	509	1108	757
# of test cases	84	50	110	76

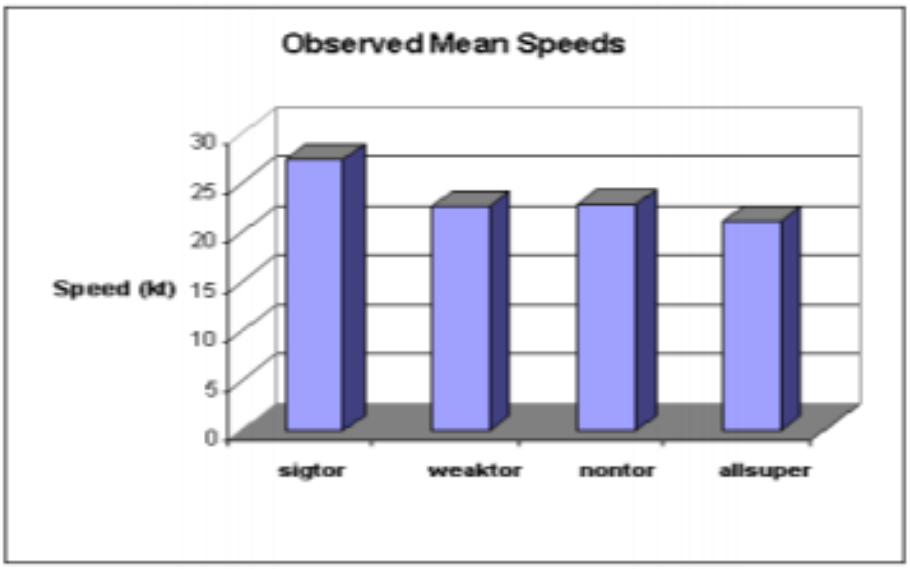


Figure 3.1. Mean observed storm speed for the following supercell (storm) groups: 56 significant tornadic storms (sigtor), 151 weak tornadic storms (weaktor), 245 nontornadic storms (nontor), and all supercells combined (allsuper). Storm speed shown is in knots. "Observed" storm motion was sampled within approximately one hour of the tornadic phase for tornadic storms, and at least an hour after convective initiation for all others. The observed motion for each supercell was computed using a distance/direction tracking component of Storm Prediction Center (SPC) operational software. Motions were derived from real-time WSR-88D reflectivity displays, following the storm centroids through a series of volume scans lasting > 30 minutes per storm. Figure and description from Edwards et al. (2002).

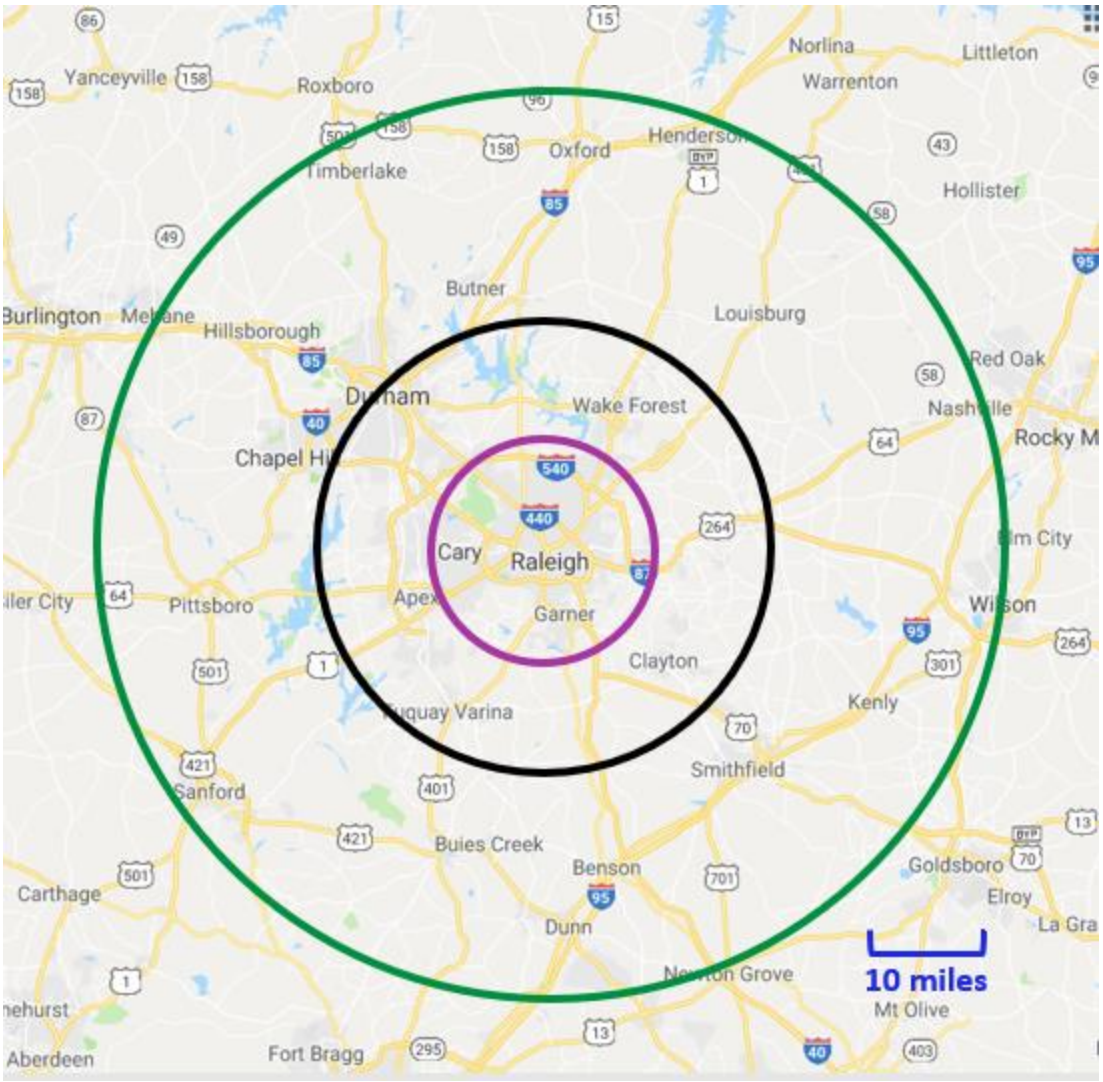


Figure 3.2. Geographic map to visualize the sample areas used to calculate maximum and minimum values of predictors for the new training datasets (i.e., NARR 5x5, NARR 3x3, RAP 5x5, and RAP 3x3). Colored circles centering around Raleigh represent the sample radiuses of 64.8 km (40.2 miles) in green, 32.4 km (20.1 miles) in black, and 16.2 km (10 miles) in purple. Google Maps background. Because the sample radiuses decide on which grid boxes to select in the gridded model data, the actual sample areas will be larger than the area of the circle.

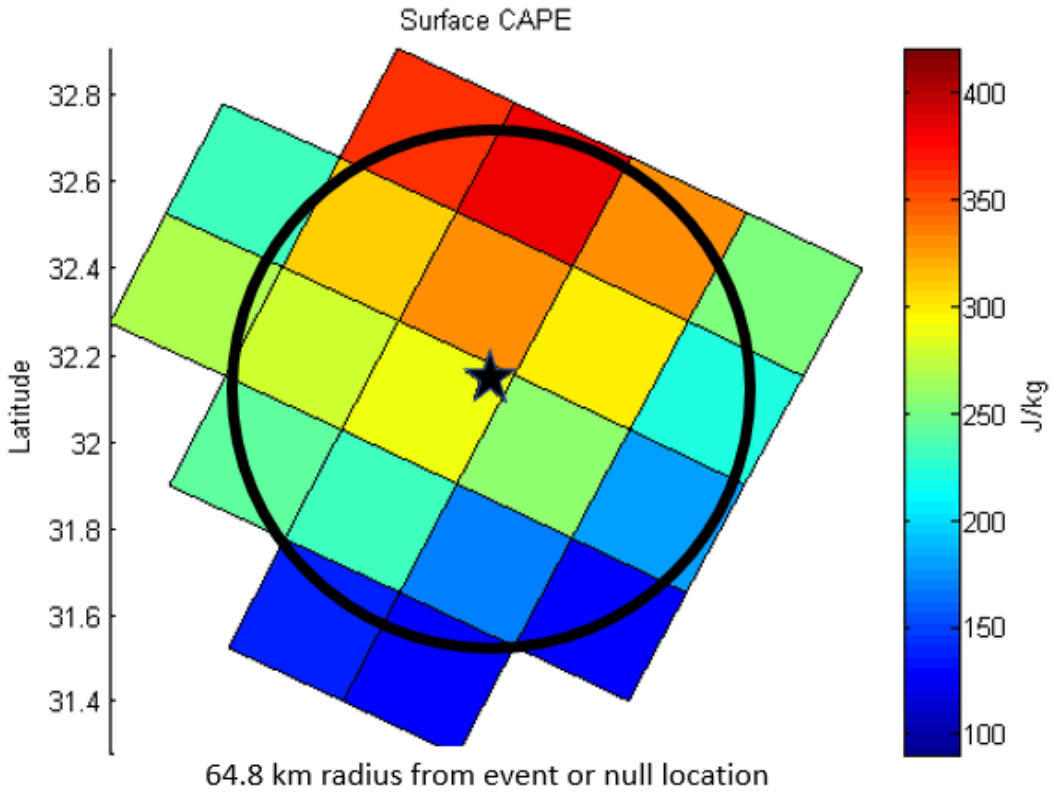


Figure 3.3. Schematic of how predictor values are calculated, using an example in the NARR 5x5 training dataset. The case located is indicated with a black star. The circle must touch the centroid of the grid box for the grid box to be included, which is why it is called a 5x5 equivalent, since usually less than 25 boxes will be included in the calculation. This allows for an understandable sample area that can be reported with the prediction (probability). This is also an effort to collect data closer to the case location and less from the environment further from the storm, especially since the maximum and minimum values are selected for the predictor values. In this example, the maximum value of the predictor (Surface CAPE) is 400 J/kg (darkest red) and the minimum value is 150 J/kg (darkest blue).

93.4352	76.7485	71.2915	79.4481	97.8970
72.1650	48.6367	39.4330	52.7374	77.7783
60.9880	29.6088	8.0616	35.9130	67.4986
65.3003	37.7141	24.7252	42.8522	71.4306
82.7047	63.2555	56.5038	66.4797	87.6755
106.9857	92.7996	88.3596	95.0681	110.9434

Figure 3.4. Example from the NARR training dataset of how the sample area from 5x5 grid boxes (green box) could consist of less than 25 grid boxes, and the sample area from 3x3 grid boxes (black box) could consist of less than 9 grid boxes. The distance from the center of the grid box to the case location is shown (in km) for each grid box. The example case shown is one of the most extreme examples since the distance from the storm report to the closest grid box was 8.1 km; if the storm report was in the center of the grid box, all grid boxes would be chosen. The sample radius for NARR 5x5 was 64.8 km, which chose 12 grid boxes. The sample radius for NARR 3x3 was 32.4 km, which chose 3 grid boxes. This allows for a smaller sample region that still captures the local storm environment while minimizing data (influence) from outside the immediate local storm environment.

CHAPTER IV

Results from Applying Statistical Procedure to New Training Datasets

4.1. Introduction

This chapter applied the severe weather statistical procedure (SWSP) to the new training datasets created in Chapter III, to test the two hypotheses: 1) the probabilistic models will be improved by using smaller sample areas to calculate the maximum and minimum of predictor values in the training dataset, and 2) the probabilistic models will be improved by using higher resolution gridded model data (i.e., gridded model with smaller grid spacing and temporal spacing) for the predictor values in the training dataset. Because the NARR dataset was already used to produce probabilistic models in Chapter II, a third hypothesis was tested using the NARR 5x5 and NARR 3x3 models: 3) predictors that are calculated from the raw predictors in the gridded model are repeating information and are thus not essential. Cutting the NARR predictor list down to the raw predictors did not affect the comparisons between the two NARR models since they shared the same predictor list. The comparison between the NARR 5x5 and RAP 3x3 models will utilize the NARR 5x5 model from Chapter II as well.

After new training datasets were created in Chapter III (i.e., NARR 5x5, NARR 3x3, RAP 5x5, and RAP 3x3), they were processed through the SWSP using R Studio. Since these datasets do not use all 25 or 9 grid boxes, these datasets can be useful for finding predictors with a strong relationship with Y that would not show up in the in larger spatial areas. The results from main steps in the SWSP were given for each training dataset. Note that new training datasets were created (section 3.4) to address the data issues mentioned in Chapters II and III. Due to the smaller number of cases (compared to Chapter II), the datasets were not split up by

severe weather hazard type, even though this was found to produce skillful models (see section 2.6). The probabilistic model created in Chapter II using the "all" dataset (tornadoes and severe wind speeds) produced a skillful model, so it was still a worthy experiment.

Because probabilistic models were trained from 5x5 and 3x3 (grid boxes) datasets from the same gridded model, the skill of these probabilistic models were compared (i.e., NARR 5x5 versus NARR 3x3, and RAP 5x5 versus RAP 3x3) to assess the sensitivity of the sample area size (sample radius). Probabilistic models that were trained from NARR 3x3 and RAP 5x5 datasets, which have the same sample area size and maximum time difference between case time and gridded model time, were compared to assess the sensitivity of spatial (horizontal) and temporal spacing in the gridded model data. However, it was discovered that the NARR 3x3 training dataset did not have enough cases to produce a reasonable probabilistic model (section 4.2.2) so this comparison was not shown. During the NARR 3x3 evaluations (sections 4.2.2 and 4.2.3), it was found that the sample radius had a small effect on the model skill, so the NARR 5x5 was also compared to the RAP 5x5 to assess the sensitivity of spatial (horizontal) and temporal spacing in gridded model data. Probabilistic models were compared using the AUROC (overall model skill), confusion matrix (to calculate FAR and POD at the optimal cutoff of the model), and total explained variation in Y.

It is important to remember that the probabilistic model was created to tell the state of the convective environment while minimizing the repeat of predictive information. In other words, the SWSP is finding the combination of skillful predictors that brings skill to the model while also contributes the most original predictive information out of the predictor list. In this chapter, the predictor list was cut down to only the raw predictors (gridded model output) on the hypothesis that predictors that are calculated from the raw predictors are repeating information

and are thus not essential. In addition, if only raw gridded model output is needed, this would speed up probabilistic model calculations and be easier to implement in real-time forecasting. It should also be kept in mind that the NARR horizontal grid spacing was 32 km and the RAP horizontal grid spacing was 13 km, and the maximum or minimum value of each predictor was chosen within a specified radius centered over the case location, so the predictor values were representing a mesoscale environment over the specified area. Therefore, the models presented in this chapter are simplified mesoscale probabilistic models that give an interpretable probability (prediction) of a severe wind speed and/or tornado within the sample radius specified.

Table 3.3 gives details on the NARR 5x5, NARR 3x3, RAP 5x5, and RAP 3x3 training datasets, including the sample radius, maximum time difference between case time and gridded model time, number of event cases, number of null cases, number of total cases (events and nulls), and the number of cases randomly selected for the testing dataset. The sample radius and model difference time (i.e., maximum time difference between case time and gridded model time) of the datasets do not limit these probabilistic models from being skillful for forecast hours further in advance; this should be tested with forecast data (see section 5.3). As discussed in section 2.6.4, to help identify predictors that are more skillful at their minimum value compared to their maximum value, the smallest deviance with Y method was used. The deviance with Y was calculated for each predictor in both the maximum value and minimum value datasets, and was compared (similar to Table 2.8). The maximum value and minimum value datasets were combined and used as the training dataset for the SWSP (see section 2.6.4 for more details.) The SWSP results from these combined training datasets are shown below.

4.2. NARR Probabilistic Models & Discussion

4.2.1. NARR 5x5 Probabilistic Model

The top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR 5x5 maximum and minimum value "all" dataset (840 cases) is shown in Table 4.1. Using the combined dataset as the training dataset for the SWSP gave the following results. The predictors in the final model and their coefficients (i.e., Estimate) are shown in the R output:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.655e+01  5.156e+00   3.209 0.001334 **
LFTX_221_ISBY          -6.634e-02  1.881e-02  -3.527 0.000420 ***
min_HGT_221_ISBL_1000  -6.002e-03  1.744e-03  -3.442 0.000577 ***
min_U_GRD_221_ISBL_250 -2.516e-02  6.551e-03  -3.840 0.000123 ***
min_V_VEL_221_ISBL_600 -3.223e-01  6.606e-02  -4.879 1.07e-06 ***
USTM_221_HTGY           8.621e-02  1.530e-02   5.634 1.76e-08 ***
V_GRD_221_ISBL_1000     7.285e-02  2.156e-02   3.379 0.000728 ***
CLWMR_221_ISBL_650     -1.368e+03  4.700e+02  -2.910 0.003619 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1163.91  on 839  degrees of freedom
Residual deviance:  974.78  on 832  degrees of freedom
AIC: 990.78
```

When validating against the entire dataset (of 840 cases), the AUROC was 0.7639 and the misclassification error was 30.7%; there were 137 false alarms (294 true nulls) and 121 missed events (288 true events). This was a FAR of 31.8% and POD of 70.4% (at the optimal cutoff). The optimal cutoff of the model was 0.4865.

Adding up the absolute value of the coefficients of the predictors (β's) reveal how much of the outcome (Y) is explained by the predictors in the model; this does not include the intercept (β_0). The β's of the predictors added up to 0.499; the predictors in the model combined contribute 49.9% of the information needed to fully predict Y (i.e., explained variation in Y). The explained variation in Y given by each predictor in the model was the following, in order

from largest to smallest contribution to an increase in the probability of an event: 0.11452 (USTM), -0.08329 (min_V_VEL_600), -0.07074 (min_U_GRD_250), -0.06259 (min_HGT_1000), 0.05927 (V_GRD_1000), -0.05467 (LFTX), -0.05355 (CLWMR_650). The u-component (west-east) of storm motion (USTM), the minimum value of omega (vertical motion) at 600 hPa (mid-levels), and the minimum value of the u-component of the wind at the jet level (250 hPa) gave the largest contributions to Y.

In general, an area of stronger upper-level divergence (area of positive vorticity advection) at the jet-level increases upward vertical motion in the atmospheric column. A negative omega means rising vertical motion and a positive omega means subsiding (downward) vertical motion. A decreasing mixing ratio at 650 hPa (drying mid-levels) and decreasing surface-to-500 hPa lifted index (LFTX) indicate convective instability and parcel instability, which increases the potential for severe weather hazards. Convective instability can occur when dry mid-level air advects over warm and moist air in the lower levels. The mixing ratio is the ratio of the mass of water vapor to the mass of dry air. Comparing to the specific humidity, which is the ratio of the mass of water vapor to the total mass (water vapor and air), the mixing ratio will be larger values, but it also represents the amount of moisture present.

The 95% confidence interval of the model's prediction (probability) revealed an uncertainty of +/- 0.10; there is a 95% confidence that the probability (%) given by the model is +/- 10.0% of the "actual" probability. See section 2.6.2 and Table 2.10 for more details.

Overall, this model was considered mediocre; however, considering this model was produced using only raw gridded model output, there is some evidence to the hypothesis that predictors calculated from other predictors (e.g., composite predictors) in the training dataset may complicate the dataset. But there is also evidence that other predictors need to be included

in the training dataset to increase overall model skill. Based on these results and the results from Chapters II, adding divergence and moist potential temperature lapse rate predictors may increase model skill without introducing too much overlapping predictive information. Positive/negative moist potential temperature lapse rate is a good indicator of stability/instability in the environment, and positive/negative divergence is an indicator upward/downward motion over a larger area. This information is missing in the raw NARR output and is important information that can distinguish between severe and non-severe (unverified warnings) environments. Therefore, the predictor list should be chosen carefully so the most skillful predictors are included in the probabilistic models.

4.2.2. NARR 3x3 Probabilistic Model

The top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR 3x3 maximum and minimum value "all" dataset (509 cases) is shown in Table 4.2. Using the combined dataset as the training dataset for the SWSP gave the following results. The predictors in the final model and their coefficients (i.e., Estimate) are shown in the R output:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.680e+01  8.820e+00  -3.039 0.002376 **
V_GRD_221_ISBL_850  8.135e-02  1.350e-02   6.024  1.7e-09 ***
U_GRD_221_ISBL_725  3.852e-02  1.411e-02   2.730 0.006332 **
CLWMR_221_ISBL_900 -2.592e+03  6.699e+02  -3.869 0.000109 ***
V_VEL_221_ISBL_350 -4.211e-01  1.645e-01  -2.560 0.010475 *
TMP_221_ISBL_100    5.942e-02  2.765e-02   2.149 0.031647 *
ICMR_221_ISBL_100  -1.453e+05  1.172e+05  -1.240 0.215033
min_TMP_221_ISBL_200  5.787e-02  2.559e-02   2.261 0.023737 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 705.61  on 508  degrees of freedom
Residual deviance: 596.23  on 501  degrees of freedom
AIC: 612.23
```

Dropping ICMR_100 slightly raised the AIC so it was kept in the model. When validating against the entire dataset (of 509 cases), the AUROC was 0.7544 and the misclassification error was 30.0%; there were 92 false alarms (161 true nulls) and 61 missed events (195 true events). This was a FAR of 36.4% and POD of 76.2% (at the optimal cutoff). The optimal cutoff of the model was 0.4619.

The predictors in the model combined contribute 49.1% of the information needed to fully predict Y (i.e., explained variation in Y). The explained variation in Y given by each predictor in the model was the following, in order from largest to smallest contribution to an increase in the probability of an event: 0.14381 (V_GRD_850), -0.08790 (CLWMR_900), 0.06064 (U_GRD_725), 0.05899 (TMP_100), -0.05643 (V_VEL_350), 0.05037 (min_TMP_200), -0.03239 (ICMR_100). The v-component (north-south) of wind at 850 hPa (around 1.5 km), cloud mixing ratio at 900 hPa (moisture around 1 km), and u-component (west-east) of wind at 725 hPa (around 2.75 km) gave the largest contributions to an increase in an event (Y=1).

Winds around 700 hPa can be a good estimate of storm motion (Thompson et al. 2003). In general, increasing northward winds at 850 hPa are associated with a strengthening low-level jet (LLJ), which moves northward at around 850 hPa. Increasing the v-component of wind can also increase the shear of the westward or eastward moving storms cells.

The 95% confidence interval of the model's prediction (probability) revealed an uncertainty of +/- 0.134; there is a 95% confidence that the probability (%) given by the model is +/- 13.4% of the "actual" probability. See section 2.6.2 and Table 2.10 for more details. Based on the AUROC, total explained variation in Y, and model uncertainty, the NARR 5x5 model was more skillful than the NARR 3x3. However, this was most likely was due to the decreased

sample size (number of cases) in the NARR 3x3 training dataset; there were 509 cases versus 840 cases in the NARR 5x5 dataset. To check if the decrease in model skill was due to the smaller sample radius, a sensitivity test was conducted.

4.2.3. Sample Radius Sensitivity Test using NARR 3x3

To ensure that the change in model skill was not due to the sample radius being "too small," the sample radius was increased 1.5 times to 48.6 km (instead of 32.4 km). This corresponding increased sample radius still fit within the 3x3 grid boxes. Therefore, this training dataset consisted of the same cases, but the maximum and minimum values were altered due to a larger sample radius. For a fair comparison, the same predictors were used in the probabilistic model to compare the change in coefficients.

Using the combined dataset as the training dataset for the SWSP gave the following results. The predictors in the final model and their coefficients (i.e., Estimate) are shown in the R output:

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.667e+01  8.848e+00  -3.014 0.002575 **
V_GRD_221_ISBL_850   8.296e-02  1.346e-02   6.162 7.18e-10 ***
U_GRD_221_ISBL_725   3.809e-02  1.416e-02   2.690 0.007142 **
CLWMR_221_ISBL_900  -2.175e+03  5.802e+02  -3.748 0.000178 ***
V_VEL_221_ISBL_350  -5.071e-01  1.986e-01  -2.554 0.010660 *
TMP_221_ISBL_100     6.549e-02  2.732e-02   2.397 0.016512 *
ICMR_221_ISBL_100   -4.719e+04  6.209e+04  -0.760 0.447218
min_TMP_221_ISBL_200 5.128e-02  2.552e-02   2.009 0.044531 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 705.61  on 508  degrees of freedom
Residual deviance: 599.48  on 501  degrees of freedom
AIC: 615.48
```

When validating against the entire dataset (of 509 cases), the AUROC was 0.7514, which was a decrease in overall model skill. The misclassification error was 29.9%; there were 86 false alarms (167 true nulls) and 66 missed events (190 true events) at the optimal cutoff. The ROC curve was more "bumpy," so there was a slight decrease in the misclassification score at the optimal cutoff, but it was overall a decrease in model skill. The optimal cutoff of the model was 0.4782.

The predictors in the model combined contribute 47.7% of the information needed to fully predict Y (i.e., explained variation in Y). The explained variation in Y given by each predictor in the model was the following, in order from largest to smallest contribution to an increase in the probability of an event: 0.14620 (V_GRD_850), -0.08552 (CLWMR_900), 0.05952 (U_GRD_725), 0.05888 (TMP_100), -0.05529 (V_VEL_350), 0.04559 (min_TMP_200), -0.02618 (ICMR_100).

There were slight changes in the coefficients (weights) of the predictors, which decreased the model skill. Because the larger sample radius decreased the model skill, this gives evidence to the hypothesis that smaller sample areas will produce more skillful probabilistic models. So, it is possible that the NARR 3x3 model would be more skillful than the NARR 5x5 model if the sample size (number of cases) was large enough in the NARR 3x3 training dataset. Also, the change in sample radius did not alter the coefficients substantially, which indicates that the coefficient estimates are not highly sensitive to the sample radius. Therefore, having a robust training dataset (enough cases) is more important than sample area size (sample radius) for producing skillful probabilistic models. A large enough number of cases (sample size) appears to be around 850 cases; the "tornado" dataset in Chapter II included 865 cases and produced a skillful probabilistic model.

To dig deeper into this comparison with the 32.4 km sample radius, the deviance with Y for each predictor was calculated for both datasets (i.e., 3x3 with 32.4 km radius and 3x3 with 48.6 km radius). Out of the 502 predictors (both maximum value and minimum values of all predictors), only 41 predictors had the deviance vary by more than 3.0, but these predictors did not significantly change the outcome of the SWSP (Table 4.3). Therefore, a moderate change in the specified sample radius may only alter the coefficients of the predictors in the model. If the sample radius is larger than recommended, the coefficients will change slightly, which can decrease model skill. It was determined that the smaller sample radius produced similar-to-more skillful coefficients, so this smaller sample radius method (see Chapter III and Table 3.3) is recommended if there are enough cases in the training dataset to do so. The larger sample radius selected all 9 grid boxes in the 3x3 grid box area, so it also shows that if a grid box method is needed for simplicity, it can be used but with a possible decrease in overall model skill.

4.3. RAP Probabilistic Models & Discussion

4.3.1. RAP 5x5 Probabilistic Model

The top 50 most skillful predictors determined by the smallest deviance with Y using the combined RAP 5x5 maximum and minimum value “all” dataset (1108 cases) is shown in Table 4.4. Using the combined dataset as the training dataset for the SWSP gave the following results, which included 13 predictors in the model.

The predictors in the final model and their coefficients (i.e., Estimate) are shown in the R output (see Table 4.6):

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.646e+01  1.058e+01  -3.447 0.000567 ***
min_VVEL_P0_L100_GLC0_525   -1.791e-01  6.393e-02  -2.801 0.005095 **
UGRD_P0_L100_GLC0_1000       9.430e-02  3.459e-02   2.726 0.006408 **
UGRD_P0_L100_GLC0_725        5.130e-02  1.133e-02   4.526 6.01e-06 ***
CIN_P0_L1_GLC0              -7.428e-03  2.887e-03  -2.573 0.010073 *
HGT_P0_L247_GLC0            -1.067e-04  3.510e-05  -3.039 0.002377 **
VGRD_P0_2L108_GLC0_60to90hPa 5.635e-02  9.985e-03   5.644 1.66e-08 ***
PRES_P0_L1_GLC0             -1.084e-04  5.199e-05  -2.085 0.037038 *
ABSV_P0_L100_GLC0            4.550e+03  1.123e+03   4.052 5.07e-05 ***
RH_P0_L100_GLC0_200          9.137e-03  2.972e-03   3.075 0.002108 **
min_POT_P0_L7_GLC0          -1.837e-02  4.676e-03  -3.930 8.51e-05 ***
TMP_P0_L100_GLC0_1000        1.706e-01  3.085e-02   5.531 3.18e-08 ***
min_HGT_P0_L243_GLC0         2.417e-05  1.225e-05   1.973 0.048505 *
min_RH_P0_L100_GLC0_875      1.110e-02  5.595e-03   1.983 0.047346 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1533.8  on 1107  degrees of freedom
Residual deviance: 1200.6  on 1094  degrees of freedom
AIC: 1228.6
```

When validating against the entire dataset (of 1108 cases), the AUROC was 0.8022 and the misclassification error was 26.3%; there were 150 false alarms (379 true nulls) and 141 missed events (438 true events). This was a false alarm rate of 28.4% and POD of 75.6% (at the optimal cutoff). Keeping in mind that the nulls in the dataset were NWS unverified severe warnings (including all severe weather hazards; see section 2.3), this was a substantial amount of correctly predicted nulls (at the optimal cutoff). The optimal cutoff of the model was 0.5251.

The predictors in the model combined contribute 66.4% of the information needed to fully predict Y (i.e., explained variation in Y). The explained variation in Y given by each predictor in the model was the following, in order from largest to smallest contribution to an increase in the probability of an event: 0.09444 (VGRD_60to90hPa), 0.08324 (TMP_1000), 0.07343 (UGRD_725), 0.05845 (ABSV_500), -0.05357 (min_POT_L7), 0.05176 (RH_200),

0.04466 (UGRD_1000), -0.04079 (min_VVEL_525), -0.04073 (HGT_L247), -0.03617 (CIN_L1), -0.03474 (PRES_L1), 0.02977 (min_RH_875), 0.02236 (min_HGT_L243).

Compared to the NARR "all" probabilistic models in section 2.6, the RAP 5x5 model is similar is overall model skill and had a lower FAR at the optimal cutoff. The RAP 5x5 model also includes more predictors, which allowed for the model to explain more of the variation in Y.

The v-component (north-south) of the wind in the 0.5 km to 1 km, and u-component (west-east) of the wind near the surface are important contributors to the prediction of an event. It is difficult to state at what height the 1000 hPa level is located and some cases have a surface above 1000 hPa, so it may be better for the model to include winds with respect to height rather than pressure level. This can still give interpretable results, as shown in section 2.6.6. U-component (west-east) winds around 700 hPa (UGRD_725) can be a good estimate of storm motion (Thompson et al. 2003), which has a larger u-component; this was also a predictor in the NARR 3x3 model. An interesting feature of the model is the height of the equilibrium level (HGT_L247) and height of the minimum convective cloud top level (HGT_L243), which are difficult predictors to forecast quantitatively in NWP models but can be viewed in atmospheric soundings (observations).

It should be noted that the tropopause height in the RAP 5x5 analysis dataset varied greatly; the event cases range from 94-598 hPa, and the null cases range from 82-419 hPa. Therefore, it is difficult to quantify what is occurring near the tropopause when this height can lower to mid-levels in the RAP model. This also shows how the tropopause heights are lower in event cases compared to null cases, which was seen in the NARR data as well. Many of the features implied by the chosen predictors are interesting and they could be represented by using other predictors, including the same variable measured in a different way (e.g., height level

instead of pressure level), to improve usability and interpretability of the probabilistic model as a whole. This can be investigated more during future work on operational forecasting applications.

The 95% confidence interval of the model's prediction (probability) revealed an uncertainty of +/- 0.1198; there is a 95% confidence that the probability (%) given by the model is +/- 12.0% of the "actual" probability, but this uncertainty can be less for different prediction ranges as seen in Table 2.10.

4.3.2. RAP 3x3 Probabilistic Model

The top 50 most skillful predictors determined by the smallest deviance with Y using the combined RAP 3x3 maximum and minimum value "all" dataset (757 cases) is shown in Table 4.5. Using the combined dataset as the training dataset for the SWSP gave the following results. The predictors in the final model and their coefficients (i.e., Estimate) are shown in the R output:

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.326e+01  8.707e+00  -3.819 0.000134 ***
UGRD_P0_L108_GLC0_150to180hPa  6.767e-02  1.072e-02   6.312 2.75e-10 ***
min_ABSV_P0_L100_GLC0          6.199e+03  1.395e+03   4.445 8.78e-06 ***
TMP_P0_L100_GLC0_1000          1.157e-01  3.060e-02   3.781 0.000156 ***
RH_P0_L100_GLC0_150            1.684e-02  4.133e-03   4.076 4.59e-05 ***
min_VGRD_P0_L100_GLC0_500      2.434e-02  1.026e-02   2.372 0.017701 *
VVEL_P0_L100_GLC0_100          1.748e+00  1.023e+00   1.709 0.087500 .
VGRD_P0_2L108_GLC0_60to90hPa   4.960e-02  1.272e-02   3.900 9.63e-05 ***
min_POT_P0_L7_GLC0            -1.099e-02  4.809e-03  -2.285 0.022315 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1044.50  on 756  degrees of freedom
Residual deviance:  837.73  on 748  degrees of freedom
AIC: 855.73
```

When validating against the entire dataset (of 757 cases), the AUROC was 0.7932 and the misclassification error was 27.5%; there were 131 false alarms (217 true nulls) and 77 missed

events (332 true events). This was a FAR of 37.6% and POD of 81.2% (at the optimal cutoff). The optimal cutoff of the model was 0.4944.

The predictors in the model combined contribute 52.9% of the information needed to fully predict Y (i.e., explained variation in Y). The explained variation in Y given by each predictor in the model was the following in order from largest to smallest contribution to an increase in the probability of an event: 0.12062 (UGRD_150to180hPa), 0.08823 (VGRD_60to90hPa), 0.07886 (RH_150), 0.07356 (min_ABSV_500), 0.05872 (TMP_1000), 0.05121 (min_VGRD_500), -0.03170 (min_POT_L7), 0.02634 (VVEL_100).

Absolute vorticity at 500 hPa and near-surface winds were chosen again in this model. The u-component (west-east) wind (UGRD_150to180hPa) right above the v-component (north-south) wind (VGRD_60to90hPa) in the 0-1.5 km above ground level could represent wind shear in the model. This was also seen in the NARR "tornado" probabilistic model (section 2.6.6). Surface temperature was chosen in this model and NARR models, which can be an important contributor of severe convective environments.

The 95% confidence interval of the model's prediction (probability) revealed an uncertainty of +/- 0.1159; there is a 95% confidence that the probability (%) given by the model is +/- 11.6% of the "actual" probability.

4.4. Model Comparisons & Summary

Based on the validation results, the NARR probabilistic models were not skillful when using a simplistic predictor list. More predictive information is needed to produce skillful models. The RAP datasets included additional predictors compared to the raw NARR output, and these extra predictors were chosen by the SWSP, which produced subsequent probabilistic

models that were more skillful. Overall, the NARR models from Chapter II had the highest overall skill compared to all probabilistic models created. Therefore, the hypothesis that a simple predictor list should be used was not validated. However, the need for removing composite parameters from predictor list should still be considered. A predictor with a smaller deviance in Y may be skillful because it combines predictive information from more than one predictor, the predictors used to calculate the parameter. The composite parameter was chosen by the SWSP because it has a smaller deviance in Y, but it usually does not provide much skill to the model as seen in section 2.6. Composite parameters were not included in the RAP predictor list and these models had overall skill similar to the Chapter II NARR models.

As seen with the NARR 3x3 and RAP 3x3, it appears that a sample radius covering 3x3 grid boxes or less does not do a good job at predictor selection; the 5x5 models appear to be more reasonable at evaluating the convective environments. However, the coefficients can be more skillful when the smaller sample radius is used to estimate the β's. Because the RAP model output was at a higher resolution, the predictors that were chosen focused on different aspects of the atmospheric column compared to the chosen predictors in the NARR model. Because of the different predictor lists available in the NARR and RAP datasets, it is difficult to compare the probabilistic models, including the Chapter II NARR 5x5 models. However, there were trends that appeared during predictor selection in all datasets, which indicated environmental predictors of importance. For example, surface temperature, low-level wind shear below 1.5 km, and u-component (west-east) winds around 700 hPa (estimate of storm motion).

It was determined that the smaller sample radius produced similar-to-more skillful coefficients, so this smaller sample radius method (see Chapter III and Table 3.3) is recommended if there are enough cases in the training dataset to do so. The larger sample radius

selected all grid boxes in the 3x3 or 5x5 grid boxes area, so it also shows that if a grid box method is needed for simplicity, it can be used but with a possible decrease in overall model skill. Overall, the statistical considerations that were covered in Chapter III are important, but there are other reasons for lacking skill in probabilistic models. This needs to be further investigated.

Keep in mind that the specific interpretation of the probabilistic models is based on the training dataset used to produce the model. This includes the case definitions (e.g., event and null definitions), gridded model used to choose the predictors and coefficients (e.g., RAP), predictor values (e.g., maximum value), sample size (e.g., sample radius), sample time (e.g., difference between observation and model time), and data type (e.g., analysis, forecast). These specifications will also determine the spatial and temporal specifications of the prediction. That is, the probability of the severe weather hazard is valid within # km and # hours. For example, in the case of the RAP 5x5 model from section 4.3.1, the sample radius was 32.4 km (20.1 miles) and the analysis (0-hour) data was used; therefore, the probabilistic model gives the probability of a severe weather hazard (wind gust ≥ 65 knots or EF1+ tornado with this model) within 32.4 km and at the time of the event (essentially nowcasting). To include uncertainties in the probability due to the forecast data (of the gridded model) used in real-time operations, forecast data needs to be used; see sections 5.2 and 5.3. To include lead time in the prediction, the forecast hour chosen (forecast data) must give lead time after the probability has been calculated; see section 5.3. Interpretation of the probabilistic models is discussed further throughout the dissertation, including Chapter I, and sections 2.3, 2.7, 5.3, 5.4, and 6.2.

Table 4.1. Top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR 5x5 maximum and minimum value "all" dataset (840 cases). The value of each predictor was taken within a sample radius of 64.8 km. The **minimum value** predictors are bolded.

Predictor	**Deviance with Y**		**Predictor**	**Deviance with Y**
TKE_950	1069		**HGT_750**	1106
TKE_925	1076		V_GRD_775	1106
HGT_1000	1089		USTM	1107
HGT_975	1090		V_GRD_900	1107
HGT_950	1092		HGT_875	1107
HGT_1000	1094		**HGT_725**	1107
V_GRD_1000	1094		U_GRD_725	1108
HGT_925	1095		HGT_850	1108
V_GRD_850	1096		**HGT_700**	1109
HGT_900	1097		U_GRD_750	1109
TKE_950	1097		HGT_825	1110
V_GRD_825	1097		U_GRD_700	1110
HGT_975	1097		V_GRD_750	1111
V_GRD_875	1098		HGT_800	1111
HGT_875	1098		U_GRD_775	1112
HGT_850	1100		HGT_775	1113
HGT_950	1100		**HGT_650**	1113
V_GRD_800	1101		V_GRD_550	1114
HGT_825	1101		**HPBL**	1114
HGT_800	1102		HGT_750	1114
HGT_925	1103		V_GRD_500	1115
VSTM	1104		V_GRD_725	1115
HGT_775	1104		V_GRD_600	1116
HGT_900	1105		HGT_725	1116
TKE_900	1106		U_GRD_800	1116

Table 4.2. Top 50 most skillful predictors determined by the smallest deviance with Y using the combined NARR 3x3 maximum and minimum value "all" dataset (509 cases). The value of each predictor was taken within a sample radius of 32.4 km. The **minimum value** predictors are bolded.

Predictor	Deviance with Y		Predictor	Deviance with Y
TKE_950	638		HGT_975	669
TKE_950	645		**HGT_900**	670
TKE_925	650		**V_GRD_1000**	671
V_GRD_850	654		**HGT_875**	671
V_GRD_875	654		HGT_950	671
V_GRD_825	657		**HGT_850**	671
V_GRD_850	658		**V_GRD_750**	671
V_GRD_825	660		V_GRD_725	672
V_GRD_875	660		**HGT_825**	672
V_GRD_800	661		V_GRD_925	672
V_GRD_900	661		HGT_925	673
TKE_925	661		**HGT_800**	673
V_GRD_1000	662		**HPBL**	673
V_GRD_800	663		HGT_900	674
HGT_1000	665		**V_GRD_925**	674
V_GRD_775	666		HGT_875	674
HGT_975	666		**HGT_775**	674
V_GRD_900	666		V_GRD_700	674
VSTM	666		**V_GRD_725**	675
HGT_950	668		HGT_850	675
HGT_1000	668		**HGT_750**	675
V_GRD_775	668		HGT_825	676
V_GRD_750	669		TKE_900	676
HGT_925	669		HGT_725	676
VSTM	669		HGT_800	676

Table 4.3. Testing the sensitivity of sample radius on predictor skill within the NARR 3x3 combined dataset. The deviance with Y for each predictor was calculated for both datasets (i.e., 3x3 with 32.4 km radius and 3x3 with 48.6 km radius). Out of the 501 predictors, only 41 predictors varied by more than 3.0 and most were less skillful with the bigger radius (48.6 km).

Predictor	Bigger Radius Deviance with Y	Smaller Radius Deviance with Y	Difference in Deviance with Y
min_V_GRD_700	680	677	3
min_U_GRD_700	692	689	3
min_U_GRD_775	695	691	3
min_U_GRD_750	694	690	3
min_U_GRD_725	693	690	3
min_V_GRD_750	675	671	3
min_V_GRD_650	683	680	3
min_V_GRD_600	685	681	4
min_CLWMR_950	703	699	4
min_V_VEL_175	703	700	4
min_TKE_850	704	701	4
min_CLWMR_925	700	696	4
min_VSTM	673	669	4
min_V_GRD_775	672	668	4
min_V_VEL_150	698	694	4
min_CLWMR_900	690	685	5
min_V_GRD_800	668	663	5
min_TKE_875	702	698	5
min_V_GRD_950	690	685	5
min_V_GRD_825	665	660	5
min_V_GRD_975	689	684	5
min_V_GRD_850	663	658	5
HPBL_SURFACE	685	680	6
min_V_GRD_900	672	666	6
min_V_GRD_875	666	660	6
min_V_GRD_925	681	674	7
min_TKE_925	670	661	9
min_TKE_900	692	683	10
PRES_SURFACE	687	693	-6
V_GRD_1000	657	662	-5
CLWMR_1000	697	701	-3
min_V_VEL_750	688	691	-3
min_V_VEL_775	688	691	-3
min_V_VEL_800	688	692	-3
min_V_VEL_725	688	691	-3
U_GRD_1000	696	699	-3
min_V_VEL_650	688	691	-3
TKE_900	673	676	-3
min_V_VEL_700	688	691	-3
min_V_VEL_825	689	692	-3
min_V_VEL_600	688	691	-3

Table 4.4. Top 50 most skillful predictors determined by the smallest deviance with Y using the combined RAP 5x5 maximum and minimum value "all" dataset (1108 cases). The value of each predictor was taken within a sample radius of 32.4 km. The **minimum value** predictors are bolded. Note that all pressure surfaces shown are level at specified pressure difference from ground to level (e.g., 120-90 hPa AGL).

Predictor	**Deviance with Y**		**Predictor**	**Deviance with Y**
VGRD_60to90hPa	1372		UGRD_725	1445
VGRD_90to120hPa	1380		UGRD_750	1447
VGRD_875	1383		UGRD_150to180hPa	1447
VGRD_900	1386		**VGRD_30to60hPa**	1448
VGRD_30to60hPa	1389		UGRD_700	1449
VGRD_850	1389		UGRD_775	1450
VGRD_120to150hPa	1395		VGRD_700	1452
VGRD_0to30hPa	1396		**VGRD_750**	1454
VGRD_825	1398		UGRD_800	1456
VGRD_150to180hPa	1407		UGRD_675	1457
VGRD_90to120hPa	1408		**HGT_1000**	1458
VGRD_800	1410		HGT_1000	1459
VGRD_850	1412		**UGRD_725**	1460
VGRD_120to150hPa	1412		**UGRD_150to180hPa**	1460
VGRD_875	1415		**VGRD_925**	1461
VGRD_825	1419		**HGT_975**	1461
VGRD_60to90hPa	1419		**UGRD_750**	1462
VGRD_925	1420		**VGRD_725**	1462
VGRD_775	1423		UGRD_120to150hPa	1462
VGRD_150to180hPa	1425		VGRD_675	1463
VGRD_900	1427		**UGRD_700**	1464
VGRD_800	1430		HGT_975	1464
VGRD_750	1434		**VGRD_0to30hPa**	1465
VGRD_725	1442		**UGRD_775**	1465
VGRD_775	1443		**HGT_950**	1465

Table 4.5. Top 50 most skillful predictors determined by the smallest deviance with Y using the combined RAP 3x3 maximum and minimum value "all" dataset (757 cases). The value of each predictor was taken within a sample radius of 16.2 km. The **minimum value** predictors are bolded. Note that all pressure surfaces shown are level at specified pressure difference from ground to level (e.g., 120-90 hPa AGL).

Predictor	**Deviance with Y**		**Predictor**	**Deviance with Y**
VGRD_60to90hPa	933		VGRD_750	977
VGRD_90to120hPa	936		**VGRD_925**	978
VGRD_60to90hPa	943		**VGRD_750**	982
VGRD_900	944		VGRD_725	985
VGRD_90to120hPa	944		UGRD_150to180hPa	986
VGRD_875	945		**UGRD_150to180hPa**	987
VGRD_120to150hPa	947		UGRD_775	987
VGRD_850	948		UGRD_750	987
VGRD_875	951		UGRD_725	988
VGRD_850	952		**VGRD_725**	990
VGRD_120to150hPa	953		UGRD_800	991
VGRD_825	954		**UGRD_775**	991
VGRD_900	954		**UGRD_750**	991
VGRD_30to60hPa	955		UGRD_700	992
VGRD_825	956		**UGRD_725**	992
VGRD_150to180hPa	958		**HGT_1000**	993
VGRD_800	961		HGT_1000	993
VGRD_150to180hPa	962		VGRD_700	994
VGRD_0to30hPa	963		**UGRD_800**	994
VGRD_800	964		UGRD_120to150hPa	995
VGRD_30to60hPa	966		**UGRD_120to150hPa**	995
VGRD_925	968		**HGT_975**	996
VGRD_775	969		**UGRD_700**	996
VGRD_775	973		HGT_975	996
VGRD_0to30hPa	974		UGRD_675	996

Table 4.6. The predictors and coefficients that make up the probabilistic model (from page 174) are shown.

Predictor (X_i)	**Coefficient (β_n)**
(intercept)	-36.46
min_VVEL_525	-0.1791
UGRD_1000	0.09430
UGRD_725	0.05130
CIN_L1	-0.007428
HGT_L247	-0.0001067
VGRD_60to90hPa	0.05635
PRES_L1	-0.0001084
ABSV_500	4550
RH_200	0.009137
min_POT_L7	-0.01837
TMP_1000	0.1706
min_HGT_L243	0.00002417
min_RH_875	0.01110

CHAPTER V

Operational Probabilistic Modeling for the National Weather Service

5.1. Introduction

The end goal is to provide analysis and operational forecasting tools specifically for the HSLC severe events observed within SEUS, which are a high-priority forecasting challenge. Currently, one of the biggest challenges to operational forecasters is limiting false alarms (i.e., false positives) and raising probability of detection (i.e., POD) of these severe events. As such, the deliverable of this research was to provide a unique severe weather statistical procedure (SWSP) that analyzes large datasets of highly correlated meteorological data and creates probabilistic forecasting models for the NWP model of interest. To test if these probabilistic models can be used "out-of-the-box" without adjustments, the models were tested with forecast data. Therefore, the research question was: How can we make these probabilistic models operationally available and useful for the National Weather Service (NWS)? The RAP probabilistic models were chosen to test for operational utility.

5.2. Creation of Testing Dataset using NWP Forecast Data

The probabilistic models from section 4.3 were verified here for operational forecasting application by using forecasting data from an operational model as the input data into the probabilistic models. Ideally, a reforecast dataset, which is a forecast dataset produced by one consistent version of a gridded (NWP) model, should be used as the training dataset to produce the probabilistic models, but reforecast data was not available. Therefore, the RAP analysis gridded (NWP) data was used as the training dataset to produce the probabilistic models, as

shown in section 4.3, and then RAP 1-hour forecast data was used as input into the models to produce a prediction. In general, the analysis data provides a more consistent dataset than forecast data, because it is constrained by observations. Analysis data will contain the physical dependencies of the predictors minus the NWP model error (bias) and was the best type of training dataset for generating the probabilistic models. To account for model error that will be present during operational use, forecast data was used as the input data into the probabilistic models to test the predictive skill in an operational setting. Gridded forecast data that share similar characteristics to the training dataset should be used as input into probabilistic model, e.g., RAP forecast data into a RAP analysis probabilistic model.

The NAM gridded model is representative of the NARR data, with the assumption that there are negligible differences in the model structure (both are Eta models); however, the horizontal grid spacing is 12 km and 32 km, respectively, which are considerably different resolutions. Because of the inconsistencies between the NAM and NARR model data, and due to the RAP model currently being used by the NWS for severe weather prediction, the RAP probabilistic models were chosen for operational testing. For a comprehensible evaluation, only cases with 13 km RAP model data (May 30, 2012 to December 25, 2015) were chosen for testing using the 1-hour forecast data. In the RAP 5x5 case list, this included 194 event cases and 224 null cases. In real-time, the input data for the probabilistic models are 13 km RAP forecast data, not analysis data. Using the RAP forecast data will include the model errors and variability in the model forecast, which is a useful test to see how the probabilistic models will work operationally.

5.3. Results with RAP Forecast

To test with a larger case list, the RAP 5x5 model (section 4.3) was tested using the RAP 1-hour forecast dataset. The optimal cutoff of 0.5251 was used again for comparisons. When validating against the entire dataset (of 418 cases), the AUROC was 0.7959 with a substantially "bumpy" ROC curve (Figure 5.1). The misclassification error was 30.4%; there were 70 false alarms (154 true nulls) and 56 missed events (138 true events). This was a FAR of 31.3% and POD of 71.1% (at the optimal cutoff).

As stated in section 4.3.1, the 95% confidence interval of the RAP 5x5 model's prediction (probability) revealed an uncertainty of +/- 0.1198; there is a 95% confidence that the probability (%) given by the model is +/- 12.0% of the "actual" probability, but this uncertainty can be less for different prediction ranges as seen in Table 2.10. This gives the probability ranges with a 95% confidence, which can be used to communicate the model's uncertainty.

These results show that the spatial and temporal definitions of the training dataset may not limit the possible skill of probabilistic models if used in further in advance (e.g., 1-hour forecast data). Because the RAP model output may not be available for 1 hour or longer in real-time operations, the 1-hour (or 2-hour) forecast data would be used for nowcasting; therefore, additional forecast hours should be tested. It also should be noted that the RAP 5x5 training dataset included cases that were up to 30 minutes before and up to 30 minutes after the RAP model time, so the 1-hour forecast data could be similar to the 0-hour analysis data.

This probabilistic model can be stated as predicting the probability of a tornado and/or severe wind speed within a 20-mile radius (around 32.4 km). The "for use in HSLC convective environments when a warning has been issued" is a disclaimer as well; see section 2.7. In the case of the RAP 5x5 model from section 4.3.1, the sample radius was 32.4 km (20.1 miles) and

the analysis (0-hour) data was used; therefore, the probabilistic model gives the probability of a severe weather hazard (wind gust ≥ 65 knots or EF1+ tornado with this model) within 32.4 km and at the time of the event (essentially nowcasting). To include uncertainties in the probability due to the forecast data (of the gridded model) used in real-time operations, forecast data needs to be used, which was done in section 5.3 and described here. To include lead time in the prediction, the forecast hour chosen (forecast data) must give lead time after the probability has been calculated in real-time. As stated here, the RAP 1-hour forecast data would not be ready for real-time use until at least an hour later, so using the RAP 1-hour forecast data would produce the prediction at the time of the predicted event. To increase this lead time, the 2-hour, 3-hour, etc., forecast data needs to be used. Further research is needed to answer questions such as: At what forecast hour does the probabilistic model lose its advantage? How far out can we forecast with the probabilistic models? This will help determine the guidelines for use of the probabilistic models.

5.4. Operational Forecasting Considerations

Because a severe thunderstorm and/or tornado warning was issued for all null cases, it was safe to assume that all cases occurred in convective environments, which was why it is stated that the resulting probabilistic models discriminate between HSLC "convective environments." This was an important distinction for the requirements necessary to use the probabilistic models in operational forecasting – the probabilistic models should only be run during the times of forecasted HSLC convective environments when a warning was issued (which is subjective). If the forecasted convective environments are outside of the HSLC defined thresholds, the probabilistic models should be used with caution. The goal of a probabilistic

model was to minimize the false alarms. This research focused on reducing FAR, because it did not address the increase in POD which needed a larger amount of missed cases in the dataset. To develop a severe weather probabilistic model that is not just based on NWS-warned cases requires developing an additional database that includes a substantial number of NWS missed events; that database would be able to address the POD forecasting issue.

The probabilistic models provide interpretable probabilities of a severe weather hazard in a specific timeframe and spatial domain, which can provide an intuitive message during emergency communications. One problem with our current forecasting techniques is a high false alarm rate. So, the focus in this research was to create a model that predicts as many nulls in the dataset as actual nulls (i.e., low false alarm rate in the dataset). This highlights the importance of choosing the event and nulls cases carefully. How an event case and null case are defined determines what the probabilistic model is predicting. To interpret the specific probabilistic models shown in this study, keep in mind the data definitions and limitations discussed in Chapter I, and sections 2.3, 2.7, and 4.4. As discussed in Chapters I-III, the specific interpretation of the probabilistic models is based on the training dataset used to produce the model. This includes the case definitions (e.g., event and null definitions), gridded model used to choose the predictors and coefficients (e.g., RAP), predictor values (e.g., maximum value), sample size (e.g., sample radius), sample time (e.g., difference between observation and model time), and data type (e.g., analysis, forecast). These specifications will also determine the spatial and temporal specifications of the prediction. That is, the probability of the severe weather hazard is valid within # km and # hours; see section 4.4 for more details. Interpretation of the probabilistic models is discussed further throughout the dissertation, including Chapter I, and sections 2.3, 2.7, 4.4, 5.3, and 6.2.

Discussions with members of the National Weather Service, including the Storm Prediction Center, during the CSTAR NWS-NCSU workshop in April 2017 identified an urgent forecasting issue – quasi-linear convective systems (QLCS) tornadoes. There are two main types of severe convection (i.e., storms that produce severe storm reports) that occur in the Southeastern United States: QLCS (i.e., essentially a line of storm cells) and discrete cells (i.e., singular storm cells). Out of the two types, the QLCS environment is the most common HSLC severe environment, as well as an important forecasting challenge for NWS forecasters (e.g., Wheatley 2008, Smith et al. 2010, Smith et al. 2012). This is a difficult forecasting challenge because these systems can generate a tornado that cannot be seen on radar due to: close proximity to the ground (i.e., rotation located below radar beam if not near the radar), narrow width of the tornado (e.g., EF0 widths can be as small as a couple dozen meters), and/or rapid development of storms within minutes. Also, the QLCS cells typically rotate, but only produce tornadoes around 2% of the time, so sending out warnings on these storms considerably increases the false alarm rate (personal communications with NWS Science and Operations Officers Jonathan Blaes and Steve Zubrick in April 2017). Within the SEUS, most of the QLCS tornadoes occurred near the coastlines, which was also the areas with the largest tornadic FAR values (Anderson-Frey et al. 2016). This was also seen with the HSLC dataset that was evaluated in Chapter II.

Compared to discrete cells, the distribution of tornadoes produced from QLCS storms mainly occurs in lower instability environments, which includes HSLC environments (Anderson-Frey et al. 2016). QLCS tornado events should be separated from RMS (supercell) tornado events regarding predictions. Anderson-Frey et al. (2016) found that QLCS tornadoes tend to occur in environments with significantly lower values of MLCAPE compared with the RMS

tornadoes (Figure 5.2); a HSLC environment is depicted in the top left corner of the plots shown in Figure 5.2. Based on these results, separating cases by storm mode (e.g., QLCS) in the training datasets could produce more skillful probabilistic models. The storm mode was a better discriminator between hits (events) and misses (nulls) than the CAPE-shear parameter space, as shown in Figure 5.2. This should be considered when developing training datasets for probabilistic models that predict the probability of a tornado.

As noted in section 2.6.3, severe weather composite parameters currently used by the NWS forecasters (e.g., STP), only focus on supercellular environments, so additional severe weather forecasting tools that either include or solely focus on QLCS environments are needed. The QLCS forecasting issue may be mitigated with additional analysis from higher resolution probabilistic models. These higher resolution probabilistic models can be produced using training datasets consisting of predictor values that are spatially and temporally closer to the storm report (e.g., Chapter III datasets), and/or using higher resolution gridded (NWP) model data (e.g., RAP and HRRR) to calculate the predictor values.

As seen throughout weather prediction studies, an ensemble prediction usually performs better than predictions from individual members of the ensemble. For a linear regression type example, an ensemble of MOS called the Consensus MOS, i.e., CMOS, takes an average of individual MOS predictions from two or more models, and has been found to improve skill compared to the individual MOS members (Vislocky and Fritsch 1995, Baars and Mass 2005). Another approach, the weighted MOS, i.e., WMOS, uses minimum variance-estimated weights to weigh each prediction from a single MOS model to improve predictions, but this showed minimal-to-no improvement in predictive skill (Daley 1991, Baars and Mass 2005). Comparing the three approaches (i.e., individual MOS, CMOS, and WMOS), the CMOS ensemble approach,

has shown to produce the most skillful prediction (Baars and Mass 2005). Not only can ensembles help improve predictions, but they also provide forecasters the ability to distinguish between high and low uncertainty forecast cases via forecast probability distributions (e.g., Toth et al. 2001). Since the 1990s, ensemble forecasting has been an integral part of weather prediction, and ensembles are used for both global and regional NWP models. More recently, the NWS has operated ensembles consisting of different variations of the same NWP model, by either changing a single parameterization scheme and/or initial boundary conditions, to account for uncertainties in the weather predictions.

Therefore, an ensemble approach can be used with the HSLC probabilistic models to see if this would improve the predictions of the severe weather hazard. An ensemble of probabilistic model contains predictors (e.g., X_1, X_2, etc.) where each predictor is an individual probabilistic model. Specifically, the probability predicted by the probabilistic model 1 is X_1, the probability predicted by probabilistic model 2 is X_2, etc. Because these probabilistic models are considered skillful in different ways, producing probabilities from all probabilistic models can give the forecaster the option of which probabilities (models) to consider based on their forecasting concerns. Similar to CMOS, the average of the predictions from two or more probabilistic models can be calculated as well, to give an ensemble mean of the predicted probabilities. For more details on how to create and utilize an ensemble of probabilistic models, see section 6.4.

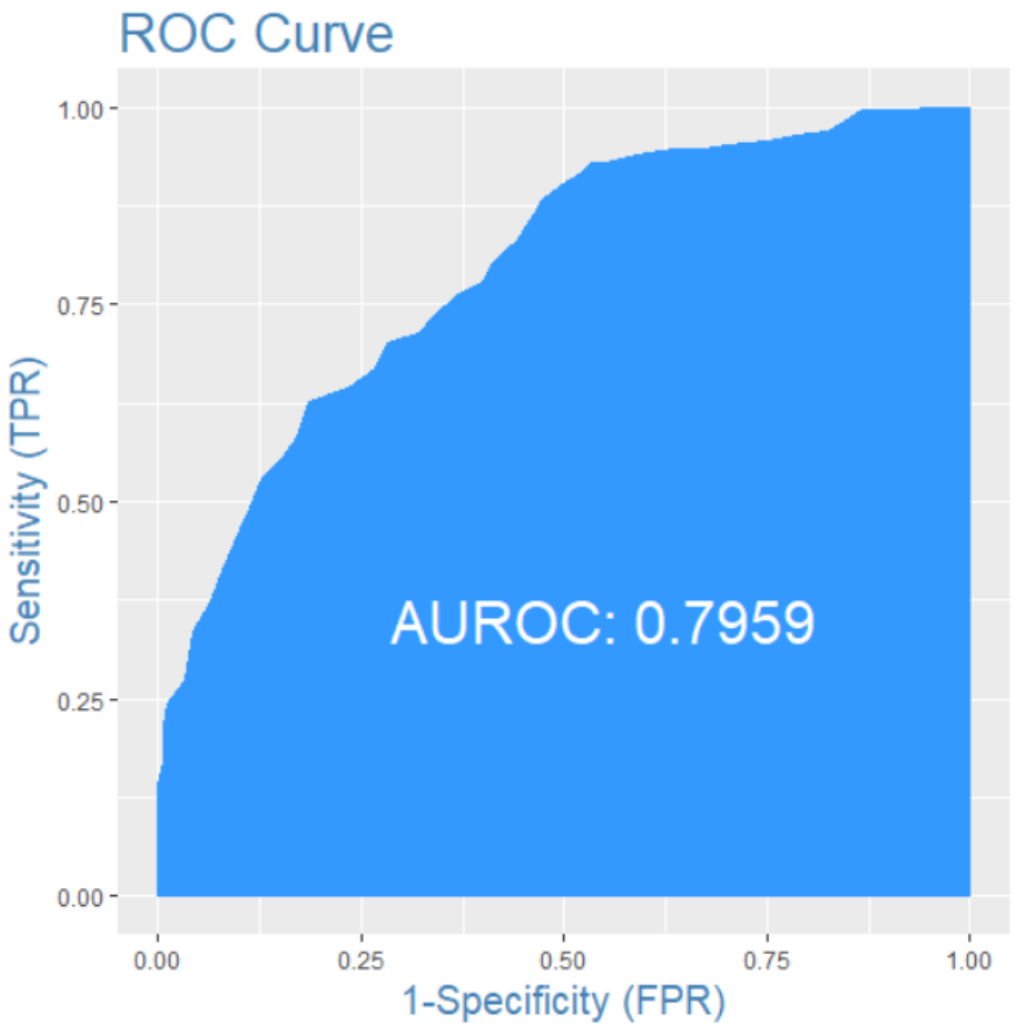


Figure 5.1. Visualization of the ROC curve and AUROC for the RAP 5x5 probabilistic model when using the RAP forecast testing dataset. The receiver operating characteristic curve (ROC curve) illustrates the skill (robustness) of a logistic regression (binary) model by plotting the true positive rate (TPR) versus the false positive rate (FPR). The TPR is also referred to as the probability of detection (POD), and the FPR is also referred to as the false alarm rate (FAR); therefore, this graphic visualizes the POD and FAR of a model.

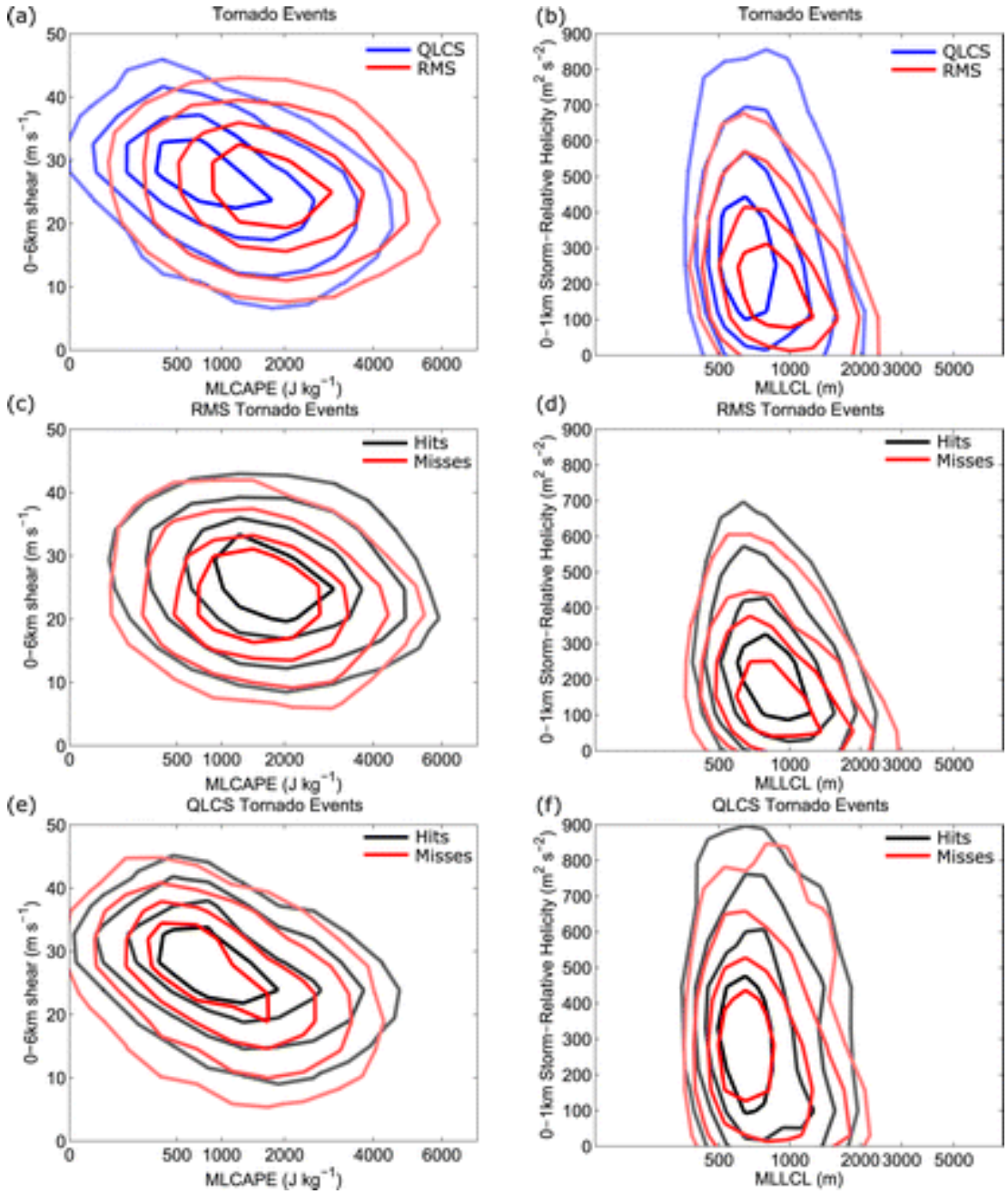


Figure 5.2. Depicting QLCS tornadoes are more common in a HSLC environment compared to RMS tornadoes, in top left graphic (a). Plots of tornado events between 2003 and 2012 in the (a) MLCAPE–SHR6 (0-6 km wind shear) and the (b) MLLCL–SRH1 (0-1 km wind shear) parameter spaces. The distribution of tornadoes from QLCS (Quasi-Linear Convective System) storms is depicted in blue, while the distribution of tornadoes from RMSs (Right-Mover Supercells) is in red. Misses (red contours; tornadoes with no prior warning) and hits (black contours; tornadoes with a warning issued ahead of touchdown) are depicted for (c),(d) RMS tornado events and (e),(f) QLCS tornado events, in the (left) MLCAPE–SHR6 and (right) MLLCL–SRH1 parameter spaces. Figure and caption details from Anderson-Frey et al. (2016).

CHAPTER VI

Discussion & Future Applications

6.1. Additional Comments

The broad goal of this research was to find ways to create more skillful severe weather forecasting models, which included separating the CAPE and shear parameter space. Through the results of this research and previous studies (e.g., Sherburn et al. 2016), it was shown that models/parameters designed for high-shear, low-CAPE (HSLC) convective environments were more skillful in predicting severe events within HSLC environments compared to other currently used severe weather composite parameters. Theoretically, the more tuned a model is to a specific environment, the more skillful the model is at predicting that environment. These specific environments could include a section of a parameter space (e.g., CAPE-shear), storm mode (e.g., QLCS), type of terrain (e.g., frictional coefficient, slope), geographical region (e.g., coastal), and/or type of severe weather hazard (e.g., tornado). For example, Coniglio et al. (2007) focused on forecasting the maintenance of warm-season QLCSs that persisted for hours. The authors took environmental variables that showed a statistically significant difference between mature and weakening Mesoscale Convective Systems (MCSs) and plugged these into a logistic regression to develop probabilistic guidance on MCS maintenance. The future of severe weather forecasting may involve hundreds of specifically tuned models instead of relying on one model to predict all severe weather in all environments. The first step is to create these specific severe weather forecasting models and then find a way to seamlessly use them with gridded (NWP) model data.

This research focused on just the HSLC quadrant of the CAPE-shear parameter space (i.e., high/low CAPE vs. high/low shear quadrants) for tornado and severe wind speeds in the

Southeastern United States (SEUS). In future work, this research can be expanded to develop probabilistic models for the other three CAPE-shear quadrants. However, a strict CAPE and shear cutoff would not operate well, so a sliding 2-D scale for a CAPE variable (e.g., SBCAPE) versus a wind shear variable (e.g., effective bulk shear) could be used. Depending on the CAPE and shear conditions, the 2-D "window" would move to the parameter space that fits the predicted CAPE-shear conditions and pull those coefficient values for each predictor in the model. The probabilistic model would have to include all the predictors necessary for all parameter spaces. Therefore, there would be a zero coefficient for the predictors that are not necessary within that "window" during that time. Currently, it is difficult to predict CAPE, so NWP models would need to improve their CAPE calculations before this method could be reasonable.

The HSLC training dataset is defined by CAPE and (deep-layer) bulk wind shear variables, thus these predictors should not be included in the probabilistic models. Therefore, the subsequent probabilistic models have a condition based on the HSLC definition. As a reminder, the HSLC environment is defined when the atmospheric environment contains 0-6 km bulk shear vector magnitude $\geq$ 18 m/s, surface-based CAPE $\leq$ 500 J/kg, and most-unstable CAPE $\leq$ 1000 J/kg (Sherburn et al. 2016). The "master" case list used by both Sherburn et al. (2016) and this research is a strictly parameter-based subset of storm reports in the SEUS, so any predictive equations created from these datasets technically are not designed to predict a severe storm environment outside of the HSLC quadrant. Hence, the SHERB(E) and MOSH(E) equations should also not be used to predict severe weather hazards outside of HSLC environments. The master case list (and subsequent datasets) consisted of only storm reports in the high-shear and low-CAPE quadrant, so the ranges of the CAPE and (deep-layer) bulk shear variables were

limited within the datasets. Because of this limited range of values within the HSLC quadrant, CAPE predictors did not significantly discriminate between severe (i.e., event) and non-severe (i.e., unverified warning, null) cases in this research. Some wind shear predictors (mostly near-surface) were skillful, but they were closely related to other predictors in the dataset, so only the simplified related predictors (winds) were used in Chapters III and IV. However, wind shear predictors were included in the predictor list in Chapter II when comparing top predictors with the Sherburn et al. (2016) top predictors and to help provide a proof of concept for the SWSP.

Additionally, there can be geographical variations in CAPE and wind shear within the same event (i.e., geographical region where storm cells are propagating during an event case), which can help explain why some areas within the convective environment are severe while nearby areas are non-severe (unverified warnings). This is one reason why sampling the convective environments in smaller sample areas can produce more skillful probabilistic models; however, NWP models may not simulate the severe environment in the same location or at the same time as observations, so this can be difficult even when using analysis data. There is also no perfect CAPE calculation, and there are errors and inconsistencies in CAPE values between NWP models. Virtual temperature corrections are used for calculating CAPE in NWP models, which helps with consistency between models, but there can still be large errors in the CAPE values compared to the values calculated with observations. Therefore, CAPE is not a predictor that should be included in probabilistic models.

6.2. Interpretation of Probabilistic Models

As a reminder, all the predictors within the probabilistic models had significant relationships with Y, which means they significantly discriminated between severe and non-

severe (unverified warnings) HSLC convective environments. The predictors in the probabilistic model cannot be interpreted individually unless holding all other predictors in the model constant. The probabilistic model is skillful because of the combination of these predictors at the weights of each predictor.

An advantage of the logistic regression method is that the probabilities given are interpretable probabilities (in comparison to coverage probabilities, for example); see Table 2.10. Some of the ways to interpret the probabilistic model are given in Step 8 of the SWSP (section 2.6) and Table 2.13. For example, one can compute how the odds of Y changes as a function of the predictor by calculating the odds-ratios; however, that is assuming all the other predictors in the model are held constant. Because a predictor in the model is likely to increase/decrease at the same time of an increase/decrease of another predictor in the model, the odds-ratios of a predictor may not give useful information. It is best to examine the probabilistic model as one complete system (e.g., Table 2.13). The predictors in the model can be centered and scaled, so the coefficients (β) can be compared; this gives the weight of the predictor in the model. After centering and scaling, the predictor with the largest coefficient is the most influential predictor in the probabilistic model. This coefficient represents the contribution of the predictor, given that all the other predictors are already in the model; therefore, this is showing the extra contribution the predictor gives to the model. Adding up the absolute values of the coefficients of the predictors (β's) reveal how much of the outcome (Y) is explained by the model (i.e., explained variation in Y). A probabilistic model containing one predictor can give the contribution of that predictor without considering other predictors in the model. The 95% confidence interval of the model's prediction can be calculated to visualize the uncertainty in the predicted probability. This uncertainty is due to the uncertainties of each predictor. This gives the overall model

uncertainty for all predicted probabilities. For a way to provide more details on the predicted probabilities in comparison to the actual probability, see Table 2.10.

To interpret the specific probabilistic models shown in this study, keep in mind the data definitions and limitations discussed in Chapter I, and sections 2.3 and 2.7. This research focused on reducing FAR, because it did not address the increase in POD which needed a larger amount of missed cases in the dataset. To develop a severe weather probabilistic model that is not just based on NWS-warned cases requires developing an additional database that includes a substantial number of NWS missed events; that database would be able to address the POD forecasting issue. The specific interpretation of the probabilistic models is based on the training dataset used to produce the model. This includes the case definitions (e.g., event and null definitions), gridded model used to choose the predictors and coefficients (e.g., RAP), predictor values (e.g., maximum value), sample size (e.g., sample radius), sample time (e.g., difference between observation and model time), and data type (e.g., analysis, forecast). Interpretation of the probabilistic models is discussed further throughout the dissertation, including Chapter I, and sections 2.3, 2.7, 4.4, 5.3, and 5.4.

The physical interpretation of each probabilistic model can be difficult to elucidate, but can be done with further research. For example, NWP modeling studies may give more information on how predictor values are related to severe weather environments in the modeling world. There are also ongoing studies that are investigating how to physically interpret results from other machine learning algorithms. Lagerquist et al. (2019) used Convolutional Neural Networks (CNNs) to produce a probability of a tornado, but this did not give an interpretable probability. Deep learning models (e.g., CNN) can learn from raw spatial grids, and we can interpret and visualize what deep learning machine models have learned by mapping the results

back onto the same spatial grid as the inputs. Lagerquist et al. (2019) used four methods (i.e., feature maps, saliency maps, backwards optimization, and novelty detection) to interpret and visualize the deep learning model results. Novelty detection can find common characteristics of the false alarms or missed events, which can help diagnose what the model is struggling with regarding skillful predictions. Because CNNs can discover new physical knowledge of how a specific severe weather hazard forms and propagates, it can be used to identify skillful predictors which could be used in logistic regression models for an interpretable probability of a severe weather hazard. Combining the knowledge gained from the SWSP and CNN methods can broaden our understanding of severe weather development.

6.2.1. Examples of Synoptic-to-Mesoscale Environmental Controls on Tornadogenesis

Multiple studies have shown that HSLC environments are strongly associated with synoptic-to-mesoscale forcing such as upper-level troughs, surface low pressure systems, and cold fronts (e.g., Guyer et al. 2006, Sherburn et al. 2016, King 2017). Synoptic scale features include baroclinic waves, troughs, ridges, and stationary waves (e.g., lee waves which are common in mountain ranges). Mesoscale features include fronts (e.g., cold front), dry line (i.e., large gradient in moisture field seen in the dewpoint temperatures), pressure centers, Low-Level Jet (LLJ), and internal gravity waves. All these meteorological features have been theorized to affect the development and/or longevity of tornadoes. When using mesoscale data in the training datasets, these features could be represented by individual or with a combination of predictors.

These larger scale features directly affect processes on the regional scale, which creates a favorable or unfavorable environment for tornadogenesis. However, these large-scale processes can control the development, strengthen, and/or weaken smaller scale processes (e.g., energy

cascade) that occur within the same environment. For example, if a front (i.e., mesoscale boundary) propagates through the environment, it will alter the temperature, moisture, and wind fields within that region, which will affect smaller scale processes including vorticity generation and buoyancy of air parcels (instability). A mesoscale boundary can enhance the depth of the boundary layer moisture and augment the low-level horizontal vorticity via baroclinic vorticity generation along the boundary (Thompson et al. 2003). Another example is a preexisting baroclinic boundary can affect a cold pool generation or strengthening/weakening. Some studies indicate cold pool strength and location are associated with tornadogenesis; for example, tornadic supercells are favored in environments that limit cold pool production (Markowski and Richardson 2014b). The large-scale features will also generate or dissipate processes occurring on the same scale; for example, a strong LLJ (usually flowing northward from a warmer, moister region) can bring in moisture ahead of a dry line, increasing the instability in that region. With ample moisture present at the low levels (within boundary layer, not at surface), convection will occur along a dryline and can become severe, which could later lead to tornadogenesis (if other factors are present).

6.2.2. Examples of Mesoscale-to-Local Environmental Controls on Tornadogenesis

Throughout the SEUS, there are more diverse terrains and various surface matter (e.g., trees) that produce frictional forces within the atmospheric boundary layer compared to the U.S. Plains, so it is hypothesized that lower atmospheric instability is enhanced by lifting mechanisms (e.g., upslope flow) to produce a portion of the annual HSLC tornado and severe wind speed reports. For example, many studies have associated low-level boundaries with tornadic storms, and it was found, during VORTEX-95, that many tornadoes occurred near these boundaries

(Markowski et al. 1998). Storms that move along a boundary can ingest enhanced horizontal vorticity and be maintained for a longer period of time. In comparison, a storm that moves over an air mass that is too cold, deep, and stable will not support low-level mesocyclogenesis, which needs tilting and stretching of the horizontal vorticity to occur.

Based on the locations of null cases in the master case list, it appears that areas along the SEUS coast may have higher false alarm rates due to additional influences from the coastal environments. This could be related to increased friction due to surface winds moving over the sea-land interface, which can increase instability in the atmospheric boundary layer, and may trigger NWP models into predicting a chance of severe storms near the coast. For example, Lombardo and Colle (2012) studied QLCS maintenance and decay on the Atlantic coastline. They found a weaker mean surface cold pool was present for the sustaining QLCSs than the decaying QLCSs, which may favor a more long-lived system if the horizontal vorticity from this cold pool is more balanced by low-level vertical shear. The QLCSs that maintained their intensity when transitioning onto land were associated with weak low-level frontogenesis, so this predictor would not be a good indicator of storm development in these cases. Numerous observational studies have shown that organized convective systems also can be influenced by smaller bodies of water, such as lakes. This gives evidence to the suggestion that geographical regions may need to be considered when producing more specific probabilistic models.

6.3. Complexity of Predicting Events in Landfalling TC Environments

The SWSP could be used to identify significant predictors in landfalling tropical cyclone (LTC) environments and then produce models that predict the probability of a tornado. The LTC convective environment is a more complex severe weather environment compared to the HSLC

environmental study, and since these occur in the SEUS, it would be an interesting comparison. Since this dissertation study is focused on predicting severe weather hazards in the SEUS, the LTC severe convective environment should be examined in future research. Refer to Appendix A for background information on TC-spawned tornadoes and LTC tornadic environments. The TCTOR (tropical cyclone tornado) dataset, discussed in Edwards (2010), contains a list of 1,295 tornadoes associated with 76 different LTCs from 1995 to 2015. Some of the LTCs produced one tornado while others produced over 100 tornadoes, so there was a wide range of activity per LTC. There were also LTCs that weakened to tropical depression strength or weaker before producing at least one tornado, so these LTC cases may need to be removed from the case list to create a robust training dataset.

Before creating a training dataset, it is necessary to first address the complexity of the predictors that should be analyzed and tested for significance and predictability. After a preliminary investigation using the SWSP, it was determined that more needs to be done during the data collection steps compared to the HSLC dataset (i.e., other predictors should be added to the predictor list for the LTC dataset). For example, it is possible that the distance and bearing (magnitude and direction) between the expected tornadogenesis location (area of interest) and the center location of the TC (given in latitude and longitude) may be skillful predictors. The TCTOR dataset (Edwards 2010) includes these predictors, among others, by pairing up each tornado case with the associated TC using the Hurricane Database (i.e., HURDAT; Neumann et al. 1999). In operations, these variables can be calculated between the location of the TC center and the location of the area of interest. Variables regarding the TC development and propagation are forecasted by the NWS National Hurricane Center (NHC), so these predictors can be found in the NHC data archive. The NHC data archive includes values for TC: maximum intensity,

intensity at landfall, size or diameter at landfall, speed and direction, type of track, and landfall location. Because these variables are routinely forecasted by the NHC for each TC, they would be available for input into an operational probabilistic model (if needed). After the TC makes landfall, the probabilistic models used for forecasting the probability of a tornado may change and become more advanced as the lead time becomes shorter. Accurately forecasting these TC variables can be very difficult, so the resulting probabilistic models (for tornado probability) will most likely be produced within the 0-hour to 3-hour lead times. These models may include predictors used for the HSLC probabilistic models. Because this study focuses on the LTC environments over land, the predictor values in the training dataset can come from higher resolution NWP models that are used in severe weather forecasting, similar to the HSLC training datasets.

The distributions of the size, depth, and convective modes of the tornadic storms in LTC environments appear to be similar to those in HSLC environments (e.g., McCaul and Weisman 2001, Edwards et al. 2012a, Edwards et al. 2012b, Davis and Parker 2014). The storms with the highest false alarm rate are typically smaller, shallower, and non-supercell (e.g., Quasi-Linear Convective Systems) in both types of environments. It is hypothesized that the HSLC and LTC convective environments share similar tornadogenesis predictors in the SEUS and that LTC environments require additional, possibly more complicated, predictors in the probabilistic model. These comparisons can be made using results from the SWSP. However, there are significant differences when comparing TC environments that are simulated at different horizontal grid spacings, as shown in numerous previous studies, so higher resolution training datasets should be created for LTC environmental studies (e.g., Weisman et al. 1997, Gentry and Lackmann 2010). Due to the horizontal grid spacing of currently available operational NWP

models, the storm-scale features of the convective storms will not be resolved, so this study should continue focusing on the environmental predictors in the regional-to-local scale (e.g., Blank 2017). Gentry and Lackmann (2010) recommended a horizontal grid spacing of 3 km or less for operational prediction of TCs before landfall, so the effects of LTCs on the environment may be well simulated at the 3 km horizontal grid spacing, which is available with the HRRR model. When enough cases (years) are available in a convection-allowing NWP model dataset (i.e., model data with horizontal grid spacing of 4 km or less), that NWP model output should be used to create a robust training dataset for probabilistic modeling. This should be done for both the LTC and HSLC environment datasets to allow for comparisons between the two.

It is recommended that the binary response variable (Y) of the probabilistic model be defined as: (1) for an event case, which is a tornadic LTC environment (i.e., numerous tornadoes reported within a sample region), and (0) for a null case, which is a non-tornadic LTC environment (i.e., no tornadoes reported within a sample region or the surrounding sample regions). However, it is difficult to find enough cases to be labeled as a null case (0), so a reasonably sized sample region should be defined based on previous LTC tornadogenesis research. Sample regions will be the same size for both event and null cases; see Chapter III for an example. Comparable to the Chapter III datasets, the center point for the event cases will be the location of the tornado report; the TCTOR dataset (Edwards 2010) includes tornado locations for all cases. The center point for a null case will be the estimated center of a false alarm (i.e., NWS tornado warning that did not verify). A reasonable distance between sample regions should also be defined, so “independent samples” can be assumed. As previously stated, spatial independence is very difficult to achieve, so this independence assumption will have that caveat.

6.4. Future Applications of Statistical Procedure

One of the next steps in this research would consist of investigating other severe convective environments like high-shear, high-CAPE (HSHC) environments or landfalling tropical cyclone (LTC) environments, and then comparing those results to the results presented in this dissertation research. For example, scientific questions for a future LTC study could include: Can the SWSP approach on LTC environments produce skillful probabilistic models? Are the identified significant predictors the same significant predictors identified in the HSLC environments? If not, what similarities exist between the HSLC probabilistic models and LTC probabilistic models?

Investigating other severe convective environments would involve creating a different case list for the training dataset based on the convective environment of interest. A larger case list with less restrictions (e.g., high-shear, all CAPE values) could also be tried, but it has been shown in this research that more specific case lists can create skilled probabilistic models. Multiple specific probabilistic models may lead to better forecasts compared to one general model. This may involve focusing on a smaller region of the United States (e.g., Gulf Coast states). As shown in Chapter III, the case list can also be restricted based on the time of the event compared to the time of the gridded (NWP) model data. For all case lists, it is recommended that the spatial and temporal spacings of the gridded model data be as small as possible for better results, recognizing that the gridded model must properly represent the convective environment at those scales for a long enough time period.

It is possible that transforming one or more of the predictors in the training dataset would create a better probabilistic model. This would also mean that the predictor must be transformed in the same way before being inputted into the model to make a prediction. To determine if a

predictor should be transformed, the user should create a scatter plot of the predictor values versus the logit of the predicted outcome, where logit = log(probabilities/(1-probabilities)). The smoothed scatter plot shows whether the predictor is linearly associated with the outcome (response variable) in logit scale, which is an assumption of the logistic model. If a scatter plot shows that a predictor is not linear, a transformation of the predictor (e.g., X^2) may be needed. Investigating predictors that are not calculated in the NWP models will be the most difficult task but may be the gap that is missing in the predictive skill of the probabilistic models. For example, weaker deep-layer shear (above 3 km) is associated with a greater potential for a cold pool to trigger convection (Coniglio et al. 2010). Predictors that are associated with cold pools may be beneficial to quantify cold pool strength, which is associated with severe weather.

A consistent training dataset is necessary to produce robust probabilistic models. The training dataset also needs to consist of hundreds to thousands of cases, which means the gridded model needs to be consistent for many years. Because most NWP models used for weather prediction are updated every 6 to 24 months (to a different version), a reanalysis or reforecast dataset is essential to providing a large enough consistent training dataset. This brings up a vital reason to advocate for the creation of reanalysis and reforecast datasets of higher resolution operational NWP models, such as the RAP, by the weather forecasting community.

The next step of this research would be utilizing convection-allowing NWP model (CAM) data as the training dataset for the SWSP outlined in Chapter II, which can contribute to the NWS Warn-On-Forecast (WoF) initiative (Stensrud et al. 2009, NSSL 2018). The WoF initiative is seeking a new approach that can “extend warning lead time in which probabilistic hazard guidance is provided by an ensemble of forecasts from convection-resolving numerical weather prediction models” (NSSL 2018). A convection-resolving model (CAM) has a

horizontal grid spacing of 4 km or less, which allows explicit convection (i.e., cumulus parameterization is turned off). Parameterizations are simplified equations that may not represent physical processes occurring in the atmosphere well but produce a model result that is close to observations. Some parameterization schemes are also modified based on the model result (comparison to observations) and not physics, so there is a benefit to not using parameterization schemes when possible. Cumulus parametrization schemes have many issues including triggering convective initiation too early or too late and producing too much or too little convection, which need to be accurate to produce realistic storms. An example of CAM is the High-Resolution Rapid Refresh (HRRR) model, which is initialized by 3 km grids (horizontal spacing) with 3 km radar assimilation every 15 minutes, as well as hourly data assimilation from the 13 km RAP model (HRRR 2018). Using the HRRR model output for training datasets would be a logical next step since the HRRR model is one of the CAMs used by the NWS for severe weather operational forecasting.

However, a CAM still cannot resolve tornadoes, but the predictor values at this higher resolution will give more details on the convective environment within the storm-scale that is conducive to tornadogenesis. CAMs can contain substantial errors in timing and location of storm cells, so an ensemble approach is ideal. Also, traditional grid-point probabilistic forecasts are not skillful at the scale of the individual storm cell (i.e., storm-scale), so an object-based approach will need to be used. An object-based approach matches the storm cell in the CAM to the location of the observed storm report based on centroid distance and minimum distance from the storm report. Part of the NOAA WoF initiative is the 3-km NSSL Experimental WoF System for Ensembles (NEWS-e), which is a rapidly updating ensemble data assimilation and prediction system operating at spatial and temporal scales of individual storms cells (Wheatley et al. 2015).

It is nested inside the HRRR ensemble (HRRRE) and has 36 WRF ensemble members that are initialized every 30 minutes with 15-minute assimilation of observations including radar, satellite, and mesonet; these members produce the analysis but there are 18 forecast ensemble members. NEWS-e gives probabilistic guidance, but it is not an interpretable probability as shown with the logistic regression method (Table 2.10).

As stated in section 3.5, because tornado warnings and severe weather warnings are typically issued within the 0-hour to 1-hour lead times, improving the 0-hour to 1-hour severe weather forecasts can minimize false alarms. Lead times could not be examined in this dissertation study due to the limited resources available. The HRRR model outputs data every 15 minutes, so when enough years (cases) are available, this NWP model can be used to examine lead times in the 0-hour to 1-hour range. This is important because, regarding tornadoes during the latest fiscal year (2017), the national averages reported were a POD of 58%, 72% FAR and a 9-minute lead time (NOAA 2018). Since improving forecasts near the time of the event can lead to forecasting improvements at larger lead times, and the average lead time for tornado warnings is currently less than 15 minutes prior to the event, it would be worthwhile to examine the 0-hour to 1-hour lead time range.

The first version of the HRRR model started operating in October 2014 but has had updated versions of the model implemented roughly every two years, so there is not a consistent dataset from the same version of the model for longer than a two-year period. Therefore, a reanalysis and/or reforecast HRRR dataset will need to be created to utilize the HRRR model output. To create a robust training dataset, the reanalysis or reforecast dataset typically needs to consist of cases over a range of eight years or more. When enough cases are available, the HRRR model will be the next operational NWP model for creating training datasets for the SWSP and

for operating the subsequent probabilistic models. A CAM training dataset, such as the HRRR model output, will create higher resolution probabilistic models that focus on storm-scale environments, which is essential for severe weather forecasting. A training dataset that is populated from a CAM that has a small enough temporal spacing (within minutes of the reported severe weather hazard) may not need maximum and minimum values of each variable, as discussed in Chapter III; the actual predictor value calculated by the CAM within the nearest grid box to the report may be sufficient.

Using the HRRR ensemble (HRRRE) output, which is currently in the experimental stage, as input into the HRRR probabilistic ensemble would be the next application (Dowell et al. 2018). Two options for trying this are: 1) calculate a probability for each of the 40 ensemble members and plot all predicted probabilities to show the range of possible outcomes, or 2) calculate one probability using the ensemble's minimum value of the predictor for a minimum value predictor and the ensemble's maximum value of the predictor for a maximum value predictor; this means different ensemble members could be chosen for each predictor value in the probabilistic equation. The first option would be a practical way to show uncertainties in the prediction. The second option should only be used if the probabilistic models were trained using maximum and/or minimum predictor values within an area consisting of more than one gridded (NWP) model grid box, such as the Chapter III data collection methods. If the HRRRE temporal spacing was decreased to 15 minutes (like the current operational HRRR model), the current spatial resolution may allow for data collection methods that involve one grid point, e.g., higher resolution storm-centered methods that follow the center of the storm. If this were the case, the first option would be more practical. The steps for this option (1) would include:

- Plotting the probability range from the ensemble. Each ensemble member reports one probability, which will be plotted as a histogram with frequency of probability on the y-axis and probability value (from 0 to 1 with 0.1 intervals) on the x-axis.

- Reporting the probability range from the ensemble to illustrate uncertainty. For example: “the ensemble predicted a probability within the range of 0.3 to 0.5 and a median of 0.4.”

- Using the probability value with the highest frequency as the ensemble’s probabilistic forecast. The median probability (ensemble mean) also can be reported, depending on the distribution of the probabilities.

One future addition to the operational use of the probabilistic models could be a NWS forecaster getting all the predictor values for input into the probabilistic model(s) with one click (i.e., cursor readout) in AWIPS2, which is the data download and analysis software used by the NWS. This would allow the forecaster to see which predictor(s) is influencing the predicted outcome (probability) the most. The predicted probability would be given at that location through the cursor readout. These real-time predictor values could be entered manually into the probabilistic equation for a quick prediction, if needed. For example, if there is a geographical area of concern, the forecaster could click in this region and the cursor readout will list the values of each predictor; this is assuming the training dataset for the probabilistic model consisted of predictor values from one grid box (e.g., storm-centered data). Any gridded (NWP) model available to the NWS operational forecasters can be used to produce these predictor values; however, which NWP model output to use would be based on the training dataset used to train the probabilistic model. The predictors in the probabilistic model can be pre-selected in the software so only those predictor values show in the cursor readout, which is currently being done for severe weather predictors of interest. Because the probabilistic models will consist of less

than fifteen predictors, this can reasonably be done. This cursor readout capability is available in the new NSEA Application as of late 2017 (personal communications with NWS Science and Operations Officers Jonathan Blaes and Steve Zubrick in April 2017).

As discussed in sections 5.4 and 6.2, this research focused on reducing FAR because the dataset used in this study only included cases in which a severe thunderstorm or tornado warning was issued in an HSLC environment. The model developed in this study can be extended to predict un-warned tornadic cases in HSLC environment in the future when such cases with sufficiently large sample size become well-documented and available. Finally, it should be noted that the cases identified as FAR may also contain uncertainty as tornadic events could occasionally go undetected. Such uncertainty can be reduced by improving the case dataset in future studies.

REFERENCES

Ahijevych, D., J.O. Pinto, J.K. Williams, and M. Steiner, 2016: Probabilistic Forecasts of Mesoscale Convective System Initiation Using the Random Forest Data Mining Technique. *Wea. Forecasting*, **31**, 581–599.

Alfaro-Cordoba, M., 2017: Variable Selection Methods with Applications to Atmospheric Sciences, Ph.D. dissertation, Dept. of Statistics, North Carolina State University, 117 pp.

Anderson-Frey, A.K., Y.P. Richardson, A.R. Dean, R.L. Thompson, and B.T. Smith, 2016: Investigation of Near-Storm Environments for Tornado Events and Warnings. *Wea. Forecasting*, **31**, 1771–1790.

Baars, J.A. and C.F. Mass, 2005: Performance of National Weather Service Forecasts Compared to Operational, Consensus, and Weighted Model Output Statistics. *Wea. Forecasting*, **20**, 1034–1047.

Benjamin, S.G., D. Dévényi, S.S. Weygandt, K.J. Brundage, J.M. Brown, G.A. Grell, D. Kim, B.E. Schwartz, T.G. Smirnova, T.L. Smith, and G.S. Manikin, 2004: An Hourly Assimilation–Forecast Cycle: The RUC. *Mon. Wea. Rev*., **132**, 495–518.

Benjamin, S.G., S.S. Weygandt, J.M. Brown, M. Hu, C.R. Alexander, T.G. Smirnova, J.B. Olson, E.P. James, D.C. Dowell, G.A. Grell, H. Lin, S.E. Peckham, T.L. Smith, W.R. Moninger, J.S. Kenyon, and G.S. Manikin, 2016: A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694.

Berger, J. O., 1985: *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer Series in Statistics, Vol. XVI, Springer, 617 pp.

Blank, L.R., 2017: Operational predictability of explicit high-shear, low-CAPE convection. M.S. thesis, Dept. of Marine, Earth, and Atmospheric Sciences, North Carolina State University, 147 pp.

Brady, R.H., and E.J. Szoke, 1989: A case study of nonmesocyclone tornado development in northeast Colorado: similarities to waterspout formation. *Mon. Wea. Rev*., **117**, 843–856.

Brooks, H.E., C.A. Doswell III, and M.P. Kay, 2003: Climatological estimates of local daily tornado probability for the United States. *Wea. Forecasting*, **18**, 626–640.

Brotzge, J. and S. Erickson, 2009: NWS Tornado Warnings with Zero or Negative Lead Times. *Wea. Forecasting*, **24**, 140–154.

Brotzge, J., S. Erickson, and H. Brooks, 2011: A 5-yr climatology of tornado false alarms. *Wea. Forecasting*, **26**, 534–544.

Brotzge, J.A., S.E. Nelson, R.L. Thompson, and B.T. Smith, 2013: Tornado Probability of Detection and Lead Time as a Function of Convective Mode and Environmental Parameters. *Wea. Forecasting*, **28**, 1261–1276.

Caruso, J.M., and J.M. Davies, 2005: Tornadoes in nonmesocyclone environments with pre-existing vertical vorticity along convergence boundaries. *Electron. J. Oper. Meteor.*, **6** (4), 1–36.

Chen, Z., W. Hendrix, H. Guan, I.K Tetteh, A. Choudhary, F. Semazzi, and N.F. Samatova, 2013: Discovery of extreme events-related communities contrasting groups of physical system networks. *Data Min. Knowl. Disc.*, **27** (2).

CISL (Computational and Information Systems Laboratory), cited 2017. Yellowstone: IBM iDataPlex System. Boulder, CO: National Center for Atmospheric Research. [Available online at http://n2t.net/ark:/85065/d7wd3xhc].

CISL (Computational and Information Systems Laboratory), cited 2018. Cheyenne: HPE/SGI ICE XA System (University Community Computing). Boulder, CO: National Center for Atmospheric Research, doi:10.5065/D6RX99HX.

Coniglio, M.C., H.E. Brooks, S.J. Weiss, and S.F. Corfidi, 2007: Forecasting the Maintenance of Quasi-Linear Mesoscale Convective Systems. *Wea. Forecasting*, **22**, 556–570.

Coniglio, M.C., J.Y. Hwang, and D.J. Stensrud, 2010: Environmental Factors in the Upscale Growth and Longevity of MCSs Derived from Rapid Update Cycle Analyses. *Mon. Wea. Rev*., **138**, 3514–3539.

Cox, D.R., 1958: The regression analysis of binary sequences (with discussion). *J Roy Stat Soc B*, **20** (2), 215–242.

Daley, R., 1991: Atmospheric Data Analysis. Cambridge University Press, 457 pp.

Davis, C.A., and K.A. Emanuel, 1991: Potential vorticity diagnostics of cyclogenesis, *Mon. Weather Rev.*, **119**, 1929–1953.

Davis, J.M., and M.D. Parker, 2014: Radar climatology of tornadic and nontornadic vortices in high-shear, low-CAPE environments in the mid-Atlantic and Southeastern United States. *Wea. Forecasting*, **29**, 828–853.

Dean, A.R., and R. S. Schneider, 2012: An examination of tornado environments, events, and impacts from 2003-2012. Preprints, 26th Conf. Severe Local Storms, Nashville TN, *Amer. Meteor. Soc.*, P6.0.

Domencich, T, and D.L. McFadden, 1996: Urban Travel Demand: A Behavioral Analysis, Chapter 5, page 124. North-Holland Publishing Co., 1975. Reprinted 1996.

Doswell, C. A. III, H. E. Brooks, and M. P. Kay, 2005: Climatological Estimates of Daily Local Nontornadic Severe Thunderstorm Probability for the United States. *Wea. Forecasting*, **20**, 577–595.

Doswell, C.A., III, H.E. Brooks, and N. Dotzek, 2009: On the implementation of the enhanced Fujita scale in the USA. *Atmos. Res*., **93**, 554–563.

Dowell, D., C. Alexander, T. Alcott, and T. Ladwig, 2018: HRRR Ensemble (HRRRE) Guidance, 2018 HWT Spring Experiment. [Available online at https://rapidrefresh.noaa.gov/internal/pdfs/2018_Spring_Experiment_HRRRE_Documentation.pdf].

Edwards, R., R.L. Thompson and J.A. Hart, 2002: Verification of Supercell Motion Forecasting Techniques. Preprints, *21st Conf. Severe Local Storms*, San Antonio TX.

Edwards, R., 2010: Tropical cyclone tornado records for the modernized National Weather Service era. Preprints, *25th Conf. on Severe Local Storms*, Denver, CO, Amer. Meteor. Soc., P2.7.

Edwards, R., A.R. Dean, R.L. Thompson, and B.T. Smith, 2012a: Convective modes for significant severe thunderstorms in the contiguous United States. Part III: Tropical cyclone tornadoes. *Wea. Forecasting*, **27**, 1114–1135.

Edwards, R., A.R. Dean, R.L. Thompson, and B.T. Smith, 2012b: Nonsupercell tropical cyclone tornadoes: Documentation, classification and uncertainties. Preprints, *26th Conf. on Severe Local Storms*, Nashville TN, Amer. Meteor. Soc., 9.6.

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. *Wea. Climate. Soc.*, **5**, 221–232.

Gagne, D., A. Mcgovern, and J. Brotzge, 2009: Classification of Convective Areas Using Decision Trees. *J Atmos Ocean Technol*, **26(7)**, 1341–1353.

Gentry, M.S., and G.M. Lackmann, 2010: Sensitivity of simulated tropical cyclone structure and intensity to horizontal resolution. *Mon. Wea. Rev*., **138**, 688–704.

Gigerenzer, G., 2004: "Mindless Statistics." *J. of Socio-Economics*, **33(5)**, 587–606.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor*., **11**, 1203–1211.

Guyer, D.A., I.A. Kis, and K. Venable, 2006: Cool Season Significant (F2-F5) Tornadoes in the Gulf Coast States. Preprints, 23rd Conf. Severe Local Storms, St. Louis MO.

Herman, G.R. and R.S. Schumacher, 2018: Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev*., **146**, 1571–1600.

Ho, Tin Kam, 1995: Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995, 278–282.

Hotelling, H., 1933: Analysis of a complex of statistical variables into principal components. *J. Edu. Psych.*, **24**, 417–441 and 498–520.

HRRR (High-Resolution Rapid Refresh) data product. Earth Systems Research Laboratory (ESRL), NOAA, cited 2018. Boulder, CO [Available online at https://rapidrefresh.noaa.gov/hrrr/].

Iwabe, C.M.N., and R.P. da Rocha, 2009: An event of stratospheric air intrusion and its associated secondary surface cyclogenesis over the South Atlantic Ocean. *J. Geophys. Res*., **114**.

Johns, R.H., J.M. Davies, and P.W. Leftwich, 1993: Some wind and instability parameters associated with strong and violent tornadoes 2.: Variations in the combinations of wind and instability parameters. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards Geophysical Monograph* **79**, Amer. Geophys. Union, 583–590.

Johnson, J.T., P.L. MacKeen, A. Witt, E.D. Mitchell, G.J. Stumpf, M.D. Eilts, and K.W. Thomas, 1998: The Storm Cell Identification and Tracking Algorithm: An Enhanced WSR-88D Algorithm. *Wea. Forecasting*, **13**, 263–276.

Jolliffe, I. T. Principal Component Analysis. 2nd ed., Springer, 2002.

Kassambara, A., 2017. Machine Learning Essentials: Practical Guide in R, Edition 1. STHDA. [Available online at www.sthda.com/english].

King, J.R., 2016: Environmental conditioning of cool season, low instability thunderstorm environments in the Tennessee and Ohio Valleys and Southeastern U.S. M.S. thesis, Dept. of Marine, Earth, and Atmospheric Sciences, North Carolina State University, 111 pp.

King, J.R, M.D. Parker, K.D. Sherburn, and G.M. Lackmann, 2017: Rapid evolution of cool season, low CAPE severe thunderstorm environments. *Wea. Forecasting*, **32**, 763–779.

King, P.S.W., 1997: On the absence of population bias in the tornado climatology of southwestern Ontario. *Wea. Forecasting*, **12**, 939–946.

Lagerquist, R., A. McGovern, C. R. Homeyer, C. K. Potvin, T. Sandmael, and T. M. Smith, 2019: Development and Interpretation of Deep Learning Models for Nowcasting Convective Hazards. 99th American Meteorological Society Annual Meeting, Phoenix, AZ (oral presentation).

Lo, F., M.C. Wheeler, H. Meinke, and A. Donald, 2007: Probability Forecasts of the Onset of the North Australian Wet Season. *Mon. Wea. Rev.*, **135**, 3506–3520.

Lombardo, K.A. and B.A. Colle, 2012: Ambient Conditions Associated with the Maintenance and Decay of Quasi-Linear Convective Systems Crossing the Northeastern U.S. Coast. *Mon. Wea. Rev.*, **140**, 3805–3819.

Markowski, P.M., E.N. Rasmussen, and J.M. Straka, 1998: The occurrence of tornadoes in supercells interacting with boundaries during VORTEX-95. *Wea. Forecasting*, **13** (3), 852–859.

Markowski, P.M., and Y.P. Richardson, 2009: Tornadogenesis: Our current understanding, forecasting considerations, and questions to guide future research. *Atmos. Res.*, **93**, 3–10.

Markowski, P.M., and Y.P. Richardson, 2014(a): What we know and don't know about tornado formation. *Physics Today*, **67** (9), 26–31.

Markowski, P.M., and Y.P. Richardson, 2014(b): The influence of environmental low-level shear and cold pools on tornadogenesis: insights from idealized simulations. *J. Atmos. Sci.*, **71**, 243–275.

Marzban, C., E.D. Mitchell, and G.J. Stumpf, 1999: The Notion of "Best Predictors": An Application to Tornado Prediction. *Wea. Forecasting*, **14**, 1007–1016.

Marzban, C., and G.J. Stumpf, 1996: A Neural Network for Tornado Prediction Based on Doppler Radar-Derived Attributes. *J. Appl. Meteor.*, **35**, 617–626.

MATLAB Release 2014a, The MathWorks Inc., Natick, Massachusetts, United States.

Mecikalski, J.R., J.K. Williams, C.P. Jewett, D. Ahijevych, A. LeRoy, and J.R. Walker, 2015: Probabilistic 0–1-h Convective Initiation Nowcasts that Combine Geostationary Satellite Observations and Numerical Weather Prediction Model Data. *J. Appl. Meteor. Climatol*., **54**, 1039–1059.

McCaul, E.W. and M.L. Weisman, 2001: The Sensitivity of Simulated Supercell Structure and Intensity to Variations in the Shapes of Environmental Buoyancy and Shear Profiles. *Mon. Wea. Rev.*, **129**, 664–687.

McGovern, A., D.J. Gagne II, N. Troutman, R.A. Brown, J. Basara, and J.K. Williams, 2011: Using spatiotemporal relational random forests to improve our understanding of severe weather processes. *Stat. Anal. Data Mining*, **4**, 407–429.

Mesinger, F., G. DiMego, E. Kalnay, K. Mitchell, P.C. Shafran, W. Ebisuzaki, D. Jović, J. Woollen, E. Rogers, E.H. Berbery, M.B. Ek, Y. Fan, R. Grumbine, W. Higgins, H. Li, Y. Lin, G. Manikin, D. Parrish, and W. Shi, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360.

NARR (North American Regional Reanalysis) data, cited 2016. NCEP Reanalysis data product provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA. [Available online at https://www.esrl.noaa.gov/psd/data/gridded/data.narr.html].

Neumann, C.J., B.R. Jarvinen, C.J. McAdie, and G.R. Hammer, 1999: Tropical Cyclones of the North Atlantic Ocean, 1871 – 1998, NOAA, Silver Springs, 206 pp.

Nicholls, N., 2001: The insignificance of significance testing. *Bull. Amer. Meteor. Soc.*, **82**(5), 981–986.

NOAA (National Oceanic and Atmospheric Administration) National Centers for Environmental Information (NCEI), cited 2019: Storm Events Database. [Available online at https://www.ncdc.noaa.gov/stormevents/].

NOAA (National Oceanic and Atmospheric Administration) National Centers for Environmental Prediction (NCEP), cited 2017: Table B, Grid Identification, Master List of NCEP Storage Grids. [Available online at http://www.nco.ncep.noaa.gov/pmb/docs/on388/tableb.html].

NOAA (National Oceanic and Atmospheric Administration) National Weather Service (NWS), cited 2018: NDFD Statistical Verification Score Definitions, Heidke Skill Score. [Available online at https://www.weather.gov/mdl/verification_ndfd_public_scoredef#hss].

NOAA (National Oceanic and Atmospheric Administration), cited 2018: Tornado warnings (nation). [Available online at https://verification.nws.noaa.gov/services/gpra/NWS_GPRA_Metrics.pdf].

NSSL (National Severe Storms Laboratory), "Warn on Forecast," cited 2018. [Available online at http://www.nssl.noaa.gov/projects/wof].

R Core Team, cited 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Available online at http://www.R-project.org].

RAP (Rapid Refresh) data, cited 2018. Data product provided by the NOAA/OAR/ESRL (Earth Systems Research Laboratory), Boulder, Colorado, USA. [Available online at https://rapidrefresh.noaa.gov].

Rasmussen, E.N., 2003: Refined Supercell and Tornado Forecast Parameters. *Wea. Forecasting*, **18**, 530–535.

Rasmussen, E.N., and D.O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164.

Röpnack, A., A. Hense, C. Gebhardt, and D. Majewski, 2013: Bayesian Model Verification of NWP Ensemble Forecasts. *Mon. Wea. Rev*., **141**, 375–387.

Schneider, A., 2016: U.S. Geological Survey (USGS) National Elevation Dataset, 1 arc-second (~30-meter) horizontal resolution. [Available online at http://www.gpsvisualizer.com/elevation].

Sherburn, K.D., and M.D. Parker, 2014: Climatology and ingredients of significant severe convection in high-shear, low-CAPE environments. *Wea. Forecasting*, **29**, 854–877.

Sherburn, K.D., M.D. Parker, J.R. King, and G.M. Lackmann, 2016: Composite environments of severe and non-severe high-shear, low-CAPE convective events. *Wea. Forecasting*, **31**, 1899–1927.

Sherburn, K.D., 2018: Environments and Origins of Low-Level Vortices within High-Shear, Low-CAPE Convection. Ph.D. dissertation, Dept. of Marine, Earth, and Atmospheric Sciences, North Carolina State University, 212 pp.

Smith, B.T., J.L. Guyer, and A.R. Dean, 2008: The Climatology, Convective Mode, and Mesoscale Environment of Cool Season Severe Thunderstorms in the Ohio and Tennessee Valleys, 1995-2006. Preprints, 24th Conf. Severe Local Storms, Savannah GA.

Smith, B.T., R.L. Thompson, J.S. Grams, and C. Broyles, 2010: Climatology of convective modes for significant severe thunderstorms in the contiguous United States. Preprints, 25th Conf. Severe Local Storms, Denver, CO.

Smith, B.T., R.L. Thompson, J.S. Grams, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States, Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135.

Smith, B.T., T.E. Castellanos, A.C. Winters, C.M. Mead, A.R. Dean, and R.L. Thompson, 2013: Measured severe convective wind climatology and associated convective modes of thunderstorms in the contiguous United States, 2003–09. *Wea. Forecasting*, **28**, 229–236.

Sobash, R.A., J.S. Kain, D.R. Bright, A.R. Dean, M.C. Coniglio, and S.J. Weiss, 2011: Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. *Wea. Forecasting*, **26**, 714–728.

Stensrud, D.J., M. Xue, L.J. Wicker, K.E. Kelleher, M.P. Foster, J.T. Schaefer, R.S. Schneider, S.G. Benjamin, S.S. Weygandt, J.T. Ferree, and J.P. Tuell, 2009: Convective-Scale Warn-on-Forecast System. *Bull. Amer. Meteor. Soc*., **90**, 1487–1500.

Thompson, R.L., R. Edwards, J.A. Hart, K.L. Elmore, and P.M. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261.

Thompson, R.L., R. Edwards, and C.M. Mead, 2004: An update to the supercell composite and significant tornado parameters. Preprints, 22nd Conf. Severe Local Storms, Hyannis, MA, *Amer. Meteor. Soc*., P8.1. [Available online at https://ams.confex.com/ams/11aram22sls/techprogram/paper_82100.htm].

Thompson, R.L., B.T. Smith, J.S. Grams, A.R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. *Wea. Forecasting*, **27**, 1136–1154.

Tochimoto, E. and H. Niino, 2017: Structural and Environmental Characteristics of Extratropical Cyclones Associated with Tornado Outbreaks in the Warm Sector: An Idealized Numerical Study. *Mon. Wea. Rev.*, **145**, 117–136.

Toth, Z., Y. Zhu, and T. Marchok, 2001: The Use of Ensembles to Identify Forecasts with Small and Large Uncertainty. *Wea. Forecasting*, **16**, 463–477.

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 post event assessment and research. *Wea. Forecasting*, **21**, 408–415.

Verbout, S.M., H.E. Brooks, L.M. Leslie, and D.M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2003. *Wea. Forecasting*, **21**, 86–93.

Vislocky, R.L. and J.M. Fritsch, 1995: Improved Model Output and Statistics through Model Consensus. *Bull. Amer. Meteor. Soc*., **76**, 1157–1164.

Wakimoto, R.M., and J.W. Wilson, 1989: Non-supercell Tornadoes. *Mon. Wea. Rev*., **117**, 1113–1140.

Weisman, M.L., W.C. Skamarock, and J.B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev*., **125**, 527–548.

Wheatley, D.M., and R.J. Trapp, 2008: The effect of mesoscale heterogeneity on the genesis and structure of mesovortices within quasi-linear convective systems. *Mon. Wea. Rev*., **136**, 4220–4241.

Wheatley, D.M., K.H. Knopfmeier, T.A. Jones, and G.J. Creager, 2015: Storm-Scale Data Assimilation and Ensemble Forecasting with the NSSL Experimental Warn-on-Forecast System. Part I: Radar Data Experiments. *Wea. Forecasting*, **30**, 1795–1817.

Widen, H.M., J.B. Elsner, C. Amrine, R.B. Cruz, E. Fraza, L. Michaels, L. Migliorelli, B. Mulholland, M. Patterson, S. Strazzo, and G. Xing, 2013: Adjusted tornado probabilities. *Electronic J. Severe Storms Meteor*., **8**(7), 1–12.

Ziliak, S.T., and D.N. McCloskey, 2009: The Cult of Statistical Significance. Preprints, Joint Statistical Meetings, Section on Statistical Education, Washington, D.C., *Amer. Stat. Assoc*., 2302–2316.

APPENDIX

Brief Review of Tornadoes Associated with Tropical Cyclone Landfall

1. Introduction

This brief review may be a helpful background for continuing the work mentioned in section 6.3. The content in this chapter was previously presented as part of a preliminary written exam taken in December 2015.

Numerous studies have found tornadoes associated with tropical cyclone (TC) landfall are common; almost all TCs making landfall in the Southeastern United States (SEUS) generated at least one tornado (McCaul 1991, McCaul and Weisman 1996, Novlan and Gray 1974, Spratt et al. 1997). However, tornado frequency is variable; some hurricanes have produced no tornadoes while others produced more than 100 tornadoes (Moore and Dixon 2011). Hurricane-strength storms produced the majority, approximately 85 percent according to one study, of tornadoes associated with TCs (Weiss 1985). Tropical cyclones can generate tornadoes from two days prior to landfall to up to three days after landfall. However, statistics show that most tornadoes occurred on the day of or the day after TC landfall (Gentry 1983, McCaul 1991, Novlan and Gray 1974). Note that "landfall" refers to the center of the TC making landfall, which means the TC outer rainbands will propagate over land prior to landfall. Regarding the time of year and time of day statistics, TC-spawned tornadoes occurred most often in August and September, and during afternoon hours (Moore and Dixon 2011, Schultz and Cecil 2009).

According to McCaul (1991), there are two distinct locations of tornadogenesis within landfalling hurricanes – core (19% of cases) and outer rainband – and most tornadoes formed within the outer rainband region where convection can be strong. Tornadoes occurred mainly within the range of 200-400 km from the hurricane center but many also occurred outside of this

range, including (on rare occasions) within the eyewall of a strong, well-organized hurricane (Figure 1, McCaul 1991). A radius of 400 km from the TC center has been identified as the maximum range of tornado activity within a hurricane, which is approximately the mean radius (351.9 km) of the outermost closed isobar of the storm (Spratt et al. 1997, Kimball and Mulekar 2004, Moore and Dixon 2011).

The Gulf Coast states (i.e., Texas, Louisiana, Mississippi, Alabama and Florida) tend to have the most frequent tornadoes associated with TCs compared to the East Coast states, mainly because of their tendency to be exposed to the right-front quadrant (RFQ) of the tropical cyclone during landfall (Novlan and Gray 1974). The RFQ of a TC is the most favorable sector of a TC for tornadogenesis because the forward momentum of the cyclone allows for faster surface wind speeds in this front sector, which is also why this sector causes the most damage upon making landfall. These winds also contribute to more favorable storm-relative hodographs (i.e., largest vertical shear and helicity) for tornadogenesis in the RFQ, which is another big contributor to property damage (McCaul 1991, McCaul and Weisman 1996). Increased surface friction and cooling of the hurricane due to landfall may also increase this wind shear (Novlan and Gray 1974), which increases chances of tornadogenesis.

2. Earlier Research on TC-Spawned Tornadoes

The majority of TC-spawned tornado research focuses on the Southern and Eastern United States because tropical cyclone landfalls occur in this region almost every hurricane season since reliable records have been kept, i.e., since around 1950 (Figure 2). One of the earliest journal articles to mention a “hurricane-spawned tornado” was Gray (1919), which was a brief one-page report published in the American Meteorological Society’s Monthly Weather Review on the occurrence and damage of a tornado that occurred in South Florida. The earliest

journal articles on this research topic, which are still cited today, include (but are not limited to) Malkin and Galway (1953), Smith (1965), Hill et al. (1966), and Orton (1970). Some of these early studies were among the first to mention low-level temperature inversion, midlevel dry (and cold) air intrusions, and low-level instability as precursors to tornadogenesis in landfalling TC environments (Malkin and Galway 1953), as well as the significance of the RFQ regarding tornadogenesis (Smith 1965, Orton 1970).

There have been numerous comprehensive studies conducted on the climatology of tornadoes associated with landfalling tropical cyclones (LTCs), including Gentry (1983), Hill et al. (1966), Moore and Dixon (2011), Novlan and Gray (1974), Schultz and Cecil (2009), Verbout et al. (2007). Figure 2 lists the major climatology studies that were discussed in Edwards (2012). There are common themes throughout the TC tornado climatology, including the suggestion that there are clear distinctions between the inner (i.e., core) and the outer-region (i.e., outer rainband) tornadoes, as previously mentioned, where outer-region tornadoes are statistically more damaging (i.e., averaged higher on the Fujita scale). Climatology studies also suggest that stronger hurricanes are more likely to produce tornado outbreaks compared to weaker hurricanes, where 78% of "outbreak hurricanes" were category 2 or higher at landfall (Verbout et al. 2007).

3. Characteristics of Tornadogenesis in Landfalling TC Environments

When a TC makes landfall, it can provide the necessary conditions for tornadogenesis; for example, wind shear and atmospheric instability. As a TC propagates over land, the surface winds are quickly slowed down due to the frictional drag over the land surface, which sets up a strong vertical wind shear profile. In tropical cyclones, most of the thermal instability is near or below 3 km, which is around half the altitude of where typical midlatitude systems maximize instability, so storms cells that develop within a TC are smaller and shallower than those found

in midlatitude severe storms (McCaul 1991, NWS 2015). The tornadic storm cells (i.e., supercells) associated with TCs have an average depth of 3-3.5 km, which is almost half the depth of the typical midlatitude supercells, averaging around 6 km (McCaul and Weisman 1996, Spratt et al 1997, Verbout et al. 2007). Because vertical wind shear is strong in the lower altitudes of a landfall TC, small supercell storms are favorable. These small supercells often produce little to no lightning and may appear innocuous on radar, but they have a high likelihood of producing tornadoes due to low-level shear as well as updrafts that are concentrated in the lowest levels. (Gentry 1983, McCaul 1991, Novlan and Gray 1974, NWS 2015).

Tornadoes in a LTC environment are common due to the small, shallow supercells that develop within the strong low-tropospheric vertical wind shear and limited instability (NOAA 2008). The TC-spawned supercells showed special similarity to “low-precipitation” (LP) supercells observed in the Great Plains, except for tornadogenesis can be hindered by the lacking depression of equivalent (moist) potential temperature aloft to allow formation of very cold downdrafts (McCaul and Weisman 1996). These cold downdrafts bring in the necessary midlevel dry air, which have been found to be a precursor to tornadogenesis. Dry air intrusions may be assisted by an enhanced upper-level (200 hPa) jet streak to the north of the storm, which accompany substantially tornadic TCs, especially those in the SEUS (Cohen 2010). A common discussion point seen throughout previous studies on TC tornadogenesis is that midlevel dry air intrusion in the RFQ of a TC (or remnant TC) will destabilize the boundary layer by steepening lapses rates and enhancing surface heating. This, in turn, will enhance low-level updrafts, and baroclinic generation of vorticity through evaporative cooling in the rear flank downdraft of the mesocyclone (supercell), increasing the potential for tornadogenesis.

If the environment is favorable for supercell storms then the possibility of tornadoes must be considered even with weakening tropical systems (i.e., remnant TCs) as these systems have been previously shown to produce tornadoes after making landfall (Vescio et al. 1996). In fact, rapidly weakening TCs are more likely to produce tornadoes compared to TCs that are slowly weakening or strengthening, upon landfall, due to the development of a colder core and abundant vertical wind shear (Novlan and Gray 1974, Moore 2015). However, strong TCs tend to produce a larger number of tornadoes compared to weak TCs (Gentry 1983, Hill et al. 1966, McCaul 1991, Moore 2015, Novlan and Gray 1974), which may be because stronger TCs can weaken at greater rates upon landfall (Moore 2015). Some studies have suggested that hurricanes recurving to the northeast after landfall were also more likely to produce tornadoes than those moving westward (Novlan and Gray 1974, Smith 1965, Verbout et al. 2007). It is interesting to note that tornadoes produced in landfalling TC environments tend to have half the path length and half the path width compared to tornadoes that form in other environments (Smith 1965). However, tornadoes with higher Fujita-Scale ratings (i.e., more severe tornadoes) generally have longer path lengths and wider path widths compared to those with lower ratings (Moore and Dixon 2011).

4. Two Example Cases

To go into more depth on the characteristics of tornadogenesis in landfalling TC environments, we will briefly discuss key aspects of two tropical cyclone case studies, Hurricane Ivan (2004) and Tropical Storm Beryl (1994), which both produced multiple tornadoes throughout the SEUS, including in the state of North Carolina. The case studies have records of tornadoes that developed within the track of the tropical cyclone on the same day and/or within the following two days after the TC made landfall. Both case studies look at the mesoscale

conditions that were in place during the time of the tornadoes within the region of interest and relate these conditions to the severe storm formation.

Hurricane Ivan (2004)

Hurricane Ivan in 2004 is well-known for being one of the largest known outbreaks of TC tornadoes. This TC produced a total of 117 (possibly more) tornadoes throughout the Southeast (and Northeast) over a three-day period, with 57 tornadoes occurring on September 17, where four of these tornadoes were produced in North Carolina (NWS 2015, Vescio et al. 1996). Hurricane Ivan reached category five status before weakening in the Gulf of Mexico as it approached the United States coastline, around the panhandle of Florida. Strong low-level wind shear along the northeastern periphery of the TC (i.e., the RFQ of the TC) allowed for the development of small supercells, nicknamed minisupercells (Baker et al. 2009, Eastin and Link 2009). Baker et al. (2009) examined these minisupercells, which were embedded within the rainbands of the RFQ of the TC during the sea-to-land transition (Figure 3). Baker et al. (2009) claimed that the significant levels of convective available potential energy (CAPE), accompanied with dry air intrusion between the rainband and main convection of TC Ivan were the primary cause of the tornado outbreaks. This study noted it was plausible for a rapidly moving supercell, embedded within a TC transitioning from a water surface to a land surface, to quickly (within 10-15 minutes) experience a dramatically different, frictionally modified environmental wind profile. Baker et al. (2009) also stated that the low-level wind shear (due to the RFQ sea-to-land transition) enhances mesocyclone updraft strength and vorticity, which creates more favorable conditions for supercells and tornadoes. Schneider and Sharp (2007) found that peak updraft intensity generally occurs at low levels and the upward dynamic pressure gradient force

contributes much more (possibly three time more) to the updraft speed than does buoyancy in a tropical cyclone-induced supercell.

Midlevel dry air intrusions have played a significant role in TCs that have affected central North Carolina, including Hurricane Ivan (Figure 4). In this event, the dry continental air mass was entrained into the TC as it traveled north from the Gulf of Mexico, over the southern states, and into North Carolina. Diurnal heating and surface-based destabilization were present due to the lack of midlevel clouds. Curtis (2004) found that TCs associated with tornadoes that had midlevel dry air intrusions also had a low average lifted condensation level (LCL), more moisture in the layer just above the surface, and were very dry above 700 hPa compared to TCs that did not produce tornadoes. Schneider and Sharp (2007) examined a sounding from Ivan, which showed the presence of a low LCL, saturation at low-levels, just above the surface through 900 hPa, and dry air at heights above 700 hPa, which agreed with the Curtis (2004) study (Figure 5). Additionally, Rasmussen and Blanchard (1998) found that LCL heights below 800 meters were associated with tornadic environments 50% of the time. A lowered LCL was found to reduce the strength of the surface outflow, preventing outflow dominance and promoting storm persistence (McCaul and Cohen 2002). Also, a low-level LCL coupled with a small surface dewpoint depression (i.e., high relative humidity at the surface) was a good indicator of a near-surface moisture profile that is favorable for tornadogenesis. This was discussed by Thompson et al. (2003) who found that increased low-level relative humidity may contribute to increased buoyancy in the rear-flank downdraft, increasing the probability of tornadoes. (Synopsis from Schneider and Sharp 2007).

Tropical Storm Beryl (1994)

A landfalling tropical cyclone does not need to be hurricane-strength to contribute to tornadogenesis, and remnant TCs have also produce tornadoes. Tropical Storm Beryl in 1994 produced a total of 37 tornadoes, including five tornadoes in North Carolina (on August 17) as a remnant tropical system (NWS 2015, Vescio et al. 1996). Tropical Storm Beryl made landfall near Panama City, Florida on August 15, 1994 and weakened to a tropical depression the following day while passing over Georgia. The tropical system continued to weaken while traveling northeast towards North Carolina, the state where the remnants of the TC produced four tornadoes, two days after making landfall. Vescio et al. (1996) found that a significant amount of dry air intruded into the northeast quadrant (i.e., RFQ) of the TC, which destabilized the boundary layer by creating steeper lapses rates and enhancing surface heating due to the lack of cloudiness. There was also increased instability and stronger storm-relative helicity, which provided a favorable environment for supercell storms. Vescio et al. (1996) hypothesized that the entrained dry air increased the potential for tornadogenesis by enhancing baroclinic generation of vorticity through evaporative cooling in the rear flank downdraft. These were similar observations to those in the Hurricane Ivan studies (Vescio et al. 1996).

Although a TC develops in a barotropic environment, it still contains low-level temperature and moisture gradients that can locally enhance storm-relative helicity, creating a favorable environment for tornadogenesis. Low-level boundaries are an important contributor to tornado development and can produce tornadoes in TCs that are undergoing an extratropical transition (Schneider and Sharp 2007). In TS Beryl, most tornadoes occurred northwest of a low-level boundary, in slightly cooler surface air with backed surface winds as the remnants of the TC passed over South Carolina (McCaul 2004). Schneider and Sharp (2007) also noted the

presence of boundaries in Beryl and its overall important role in the development of tornadoes. While Schneider and Sharp (2007) were examining Hurricane Jeanne (a similar situation to Beryl), they noted that numerous supercells were translating from south to north over an east-west oriented boundary, and several tornadoes were produced at this boundary.

5. Summary

Numerous studies have found that the majority of TCs, mostly hurricane-strength TCs, produced at least one tornado after making landfall in the SEUS. Some of the precursors to tornadogenesis associated with TC landfall discussed throughout the studies were midlevel dry air intrusions, lowered LCLs, low-level boundaries (including temperature and moisture gradients), low surface dewpoint depressions, instability focused in the low levels, and strong vertical wind shear. There are other quantitative parameters that can be used as guidance to determine if a LTC environment poses a threat of producing tornadoes (e.g., Figure 6), and there are numerous more recent studies that investigate these other predictors and/or develop new severe weather parameters for LTC environments (e.g., Edwards and Thompson 2014, Onderlinde and Fuelberg 2014, Edwards et al. 2015). For example, one of the most recently published studies, Onderlinde and Fuelberg (2014), developed a composite parameter to forecast the likelihood that one or more tornadoes (associated with landfalling TCs in the Gulf of Mexico and southern Atlantic coast) will occur within a 6-hour period based on six factors: azimuth angle of the tornado report from the tropical cyclone, distance from the cyclone's center, time of day, 0–3-km wind shear, 0–3-km storm relative helicity, and 950–1000-hPa CAPE.

One of the questions of TC-spawned tornado research that is lacking sufficient answers is: What are the distinctive environmental (physical) characteristics determining whether a TC will be active (i.e., produce numerous tornadoes within a region) versus inactive (i.e., produce no

tornadoes within the same region)? This relationship can be investigated using tools such as the severe weather statistical procedure (SWSP).

6. Figures

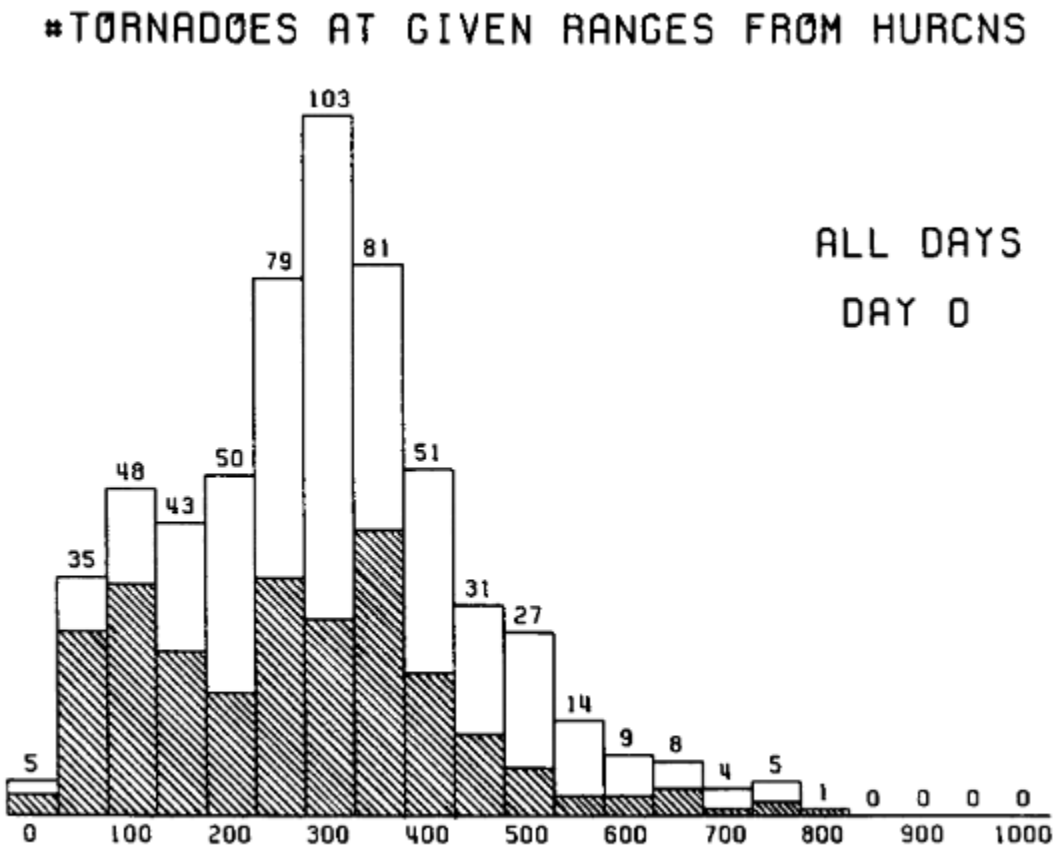


FIG. 13. Range distribution of reported hurricane tornadoes during 1948–86, for all days relative to landfall (total bar height), with tornadoes on landfall day itself indicated by hatching. Range bins are 50 km wide and are centered on the labeled values. Note distinct peak near 100-km range for landfall-day tornadoes.

Figure 1. Number of tornadoes at a given range (in km) from the hurricane center. Tornadoes occurred mainly in the 200-400 km range from the hurricane center, but many also occurred outside of this range. Figure from McCaul (1991); original figure caption included for additional information.

AUTHOR(S)	YEARS	EVENT COUNT	TC LEVELS	PLACE	INCLUSION CRITERIA
Tannehill (1944)	1811–1933	10	All	SC, FL	Unspecified
Malkin and Galway (1953)	1811–1952	22	All	U. S.	Unspecified
Wolford (1960)	1916–1957	84	All	U. S.	Unspecified
Smith (1965)	1955–1962	98	All	U. S.	"Within the cyclonic circulation"
Pearson and Sadowski (1965)	1955–1961, 1964	137	All	U. S.	Unspecified
Hill et al. (1966)	1955–1964	136	All	U. S.	"Subjectivity"
Fujita et al. (1972)	1950–1971	68	Typhoon	Japan	Unspecified
Novlan and Gray (1974)	1948–1972	373	All	U. S.	Unspecified
Gentry (1983)	1973–1981	120	All	U. S.	Unspecified, no records at $r > 350$ km
Weiss (1987)	1964–1983	397	All	U. S.	"Subjectively matched"
McCaul (1991)	1948–1986	626	All	U. S.	$r \leq 800$ km
Verbout et al. (2007)	1954–2004	1089	All	U. S.	$r \leq 400$ km, landfall ±2 days
Schultz and Cecil (2009)	1950–2007	1767	All	U. S.	$r \leq 750$ km then "inspection" for 750 km $\geq r \geq 500$ km
Belanger et al. (2009)*	1950-2008	1375	All	U. S.	Gulf landfalls, $r \leq 650$ km, only during NHC advisories
Edwards 2010	1995–2010#	1163#	All	U. S.	Meteorological analysis, no max r
Agee and Hendricks (2011)	1979–2010@	300–334@	All	FL	Pre- and post-installation of WSR-88D system
Moore and Dixon (2011)	1950–2005	734	Hurricane at landfall	U. S.	Gulf landfalls, $r \leq 400$ km, landfall ±1 day

* Available via supplemental FTP link in manuscript
# Climatology updated yearly, data complete through 2010 as of this revision
@ 1994–1995 data listed, but omitted from analyses

Figure 2. Summary of Tropical Cyclone (TC) tornado climatology in the literature. Radius r is defined from TC center. Table from Edwards (2012).

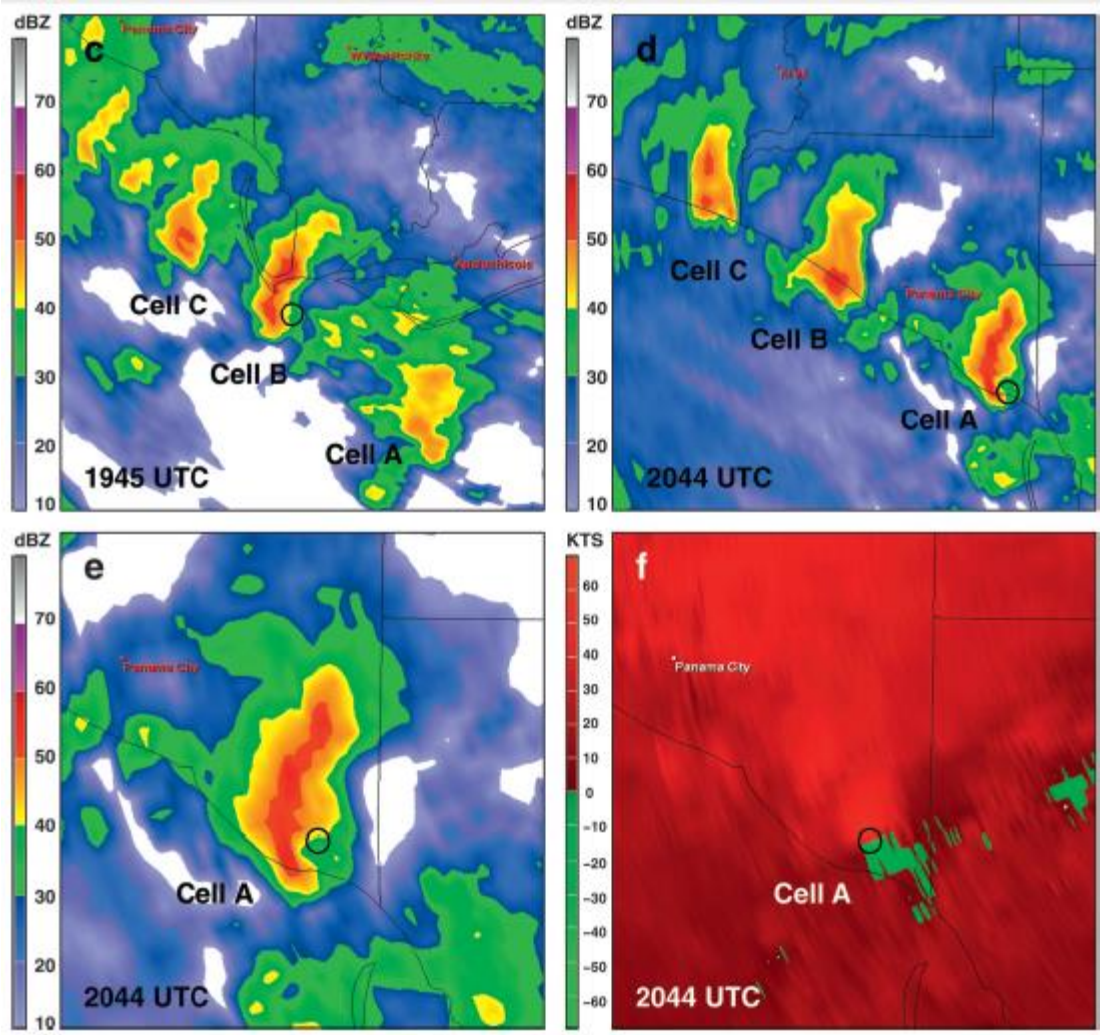


Figure 3. Radar reflectivity (c-e) and Doppler velocity (f) images of minisupercells A, B, and C viewed from the KTLH WSR-88D in Tallahassee, FL as Hurricane Ivan made landfall. This region was in the right-front quadrant of the TC during the sea to land transition. (c) detection of a mesocyclone in cell B, (d) first tornado reported from these supercells (in cell A), (e) zoomed view of cell A showing hook echo, (f) Doppler velocities at 0.5° elevation is zoomed view of cell A. Figure from Baker et al. (2009).

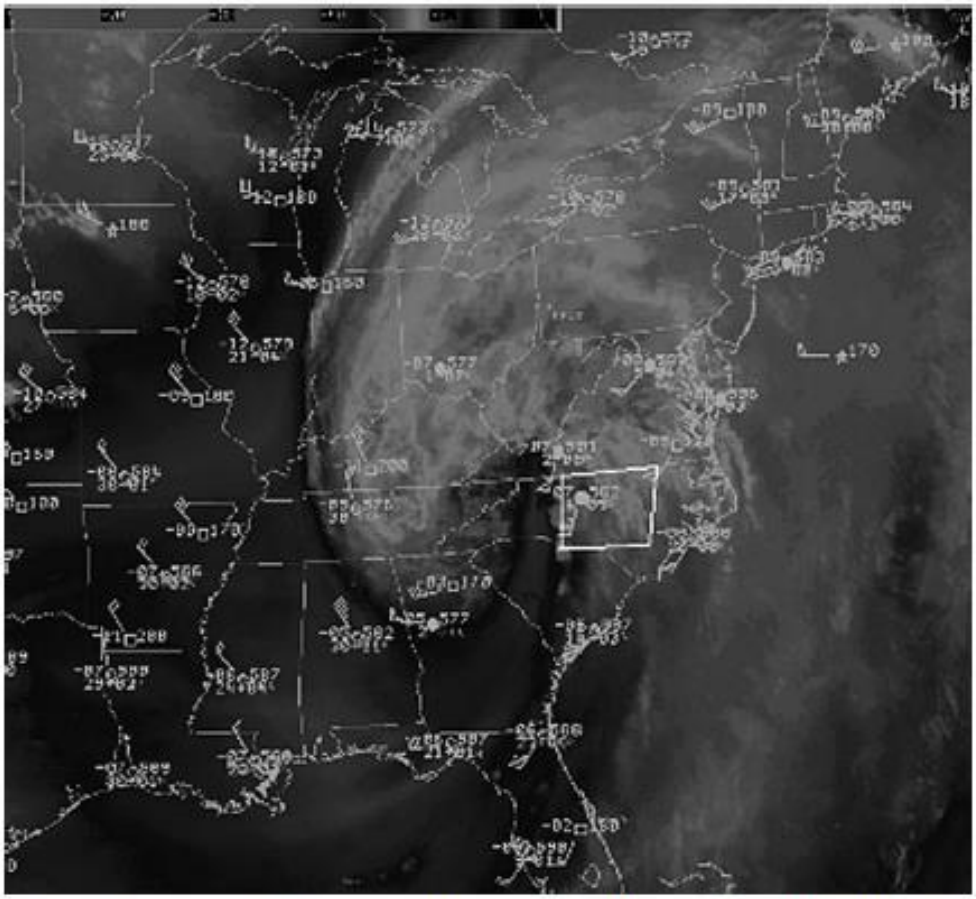


FIG. 1. Water vapor satellite imagery of the remnants of Hurricane Ivan at 1200 UTC 17 Sep 2004, with a 500-mb upper-air data plot. The center of Ivan's circulation is located about 35 mi (56 km) NW of Asheville, NC. The Raleigh CWA is indicated by the white box over central NC. A dry intrusion extends from southern GA through western NC. The 500-mb dewpoint depressions (the lower-left number in the data plots) indicate a strong moisture gradient from central NC to central TN, with dry air feeding into the circulation from AL and northern FL. At the time of this image, convection is beginning to develop along the eastern edge of the dry intrusion in NC, and a tornado was reported in the northwest portion of the Raleigh CWA about 3 h later.

Figure 4. Water vapor image of the remnants of Hurricane Ivan over North Carolina, which shows the midlevel dry air intrusion extending from southern Georgia into western NC. There is also a strong moisture gradient present from central NC to central TN. Figure from Schneider and Sharp (2007); original figure caption included for additional information.

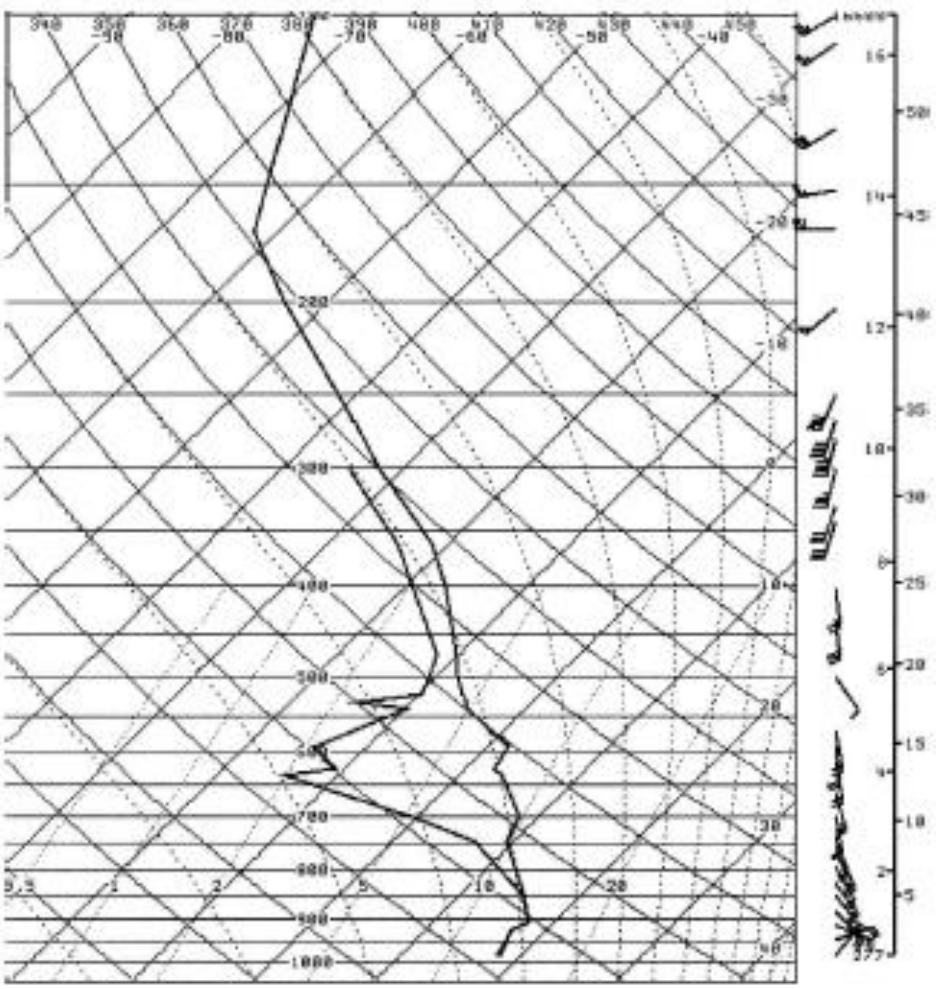


Figure 5. A skew T plot of temperature, dewpoint temperature, and winds from GSO (Greensboro, NC) as the remnants of Ivan were crossing central North Carolina. Note the low-level saturation, midlevel dry layer between 3 and 6km, and the rapid veering of winds just above the surface. Figure from Schneider and Sharp (2007).

Tornado Threat Parameters

Parameters that can be used to determine if a tropical cyclone environment poses a low or high threat of tornadoes. Adapted from McCaul (1991)

Parameter	Low Threat	High Threat
Location with respect to storm center	120 - 359 deg	1 - 120 deg
Lifted Index	> - 1	< -1
CAPE	< 500 J/kg	> 500 J/kg
0 to 3 km shear	< 39 kts (20 m/s)	> 39 kts (20 m/s)
SRH	< 100 m^2/s^2	>100 m^2/s^2
BRN	< 10 or > 50	10 to 50
850 hPa wind speed	< 30 kts (15 m/s)	> 30 kts (15 m/s)
Storm motion	< 10 mph or > 30 mph	10 to 30 mph
Axis of low level convergence over area	NO	YES

Figure 6. Tornado threat parameters that can determine if a tropical cyclone environment poses a low or high threat of tornadoes. CAPE (surface-based parcel) is the convective available potential energy. SRN refers to the 0-1 km storm-relative helicity. BRN is the Bulk Richardson number. Image from NOAA (2008).

7. References

Baker, Adam K., Matthew D. Parker, Matthew D. Eastin, 2009: Environmental Ingredients for Supercells and Tornadoes within Hurricane Ivan. *Wea. Forecasting*, **24**, 223–244.

Cohen, A.E., 2010: Synoptic-scale analysis of tornado-producing tropical cyclones along the Gulf Coast. *Natl. Weather Dig*., **34**, 99–115.

Curtis, L., 2004: Midlevel dry intrusions as a factor in tornado outbreaks associated with land-falling tropical cyclones from the Atlantic and Gulf of Mexico. *Wea. Forecasting*, **19**, 411–427.

Eastin, M.D., and M.C. Link, 2009: Miniature supercells in an offshore outer rainband of Hurricane Ivan. *Mon. Wea. Rev.*, **137**, 2081–2104.

Edwards, R., 2012: Tropical cyclone tornadoes: A review of knowledge in research and prediction. *Electronic J. Severe Storms Meteor*., **7 (6)**, 1–61.

Edwards, R., and R.L. Thompson, 2014: Reversible CAPE in tropical cyclone tornado regimes. Preprints, *27th Conf. Severe Local Storms*, Madison, WI, P88.

Edwards, R., B.T. Smith, R.L. Thompson, and A.R. Dean, 2015: Analyses of radar rotational velocities and environmental parameters for tornadic supercells in tropical cyclones. Preprints, *37th Conf. Radar Meteorology*, Norman, OK, 5A.3.

Gentry, R. C., 1983: Genesis of tornadoes associated with hurricanes. *Mon. Wea. Rev.*, **111**, 1793–1805.

Gray, R.W., 1919: A tornado within a hurricane area. *Mon. Wea. Rev.*, 639.

Hill, E.L., Malkin W., and W.A. Schulz, 1966: Tornadoes associated with cyclones of tropical origin−practical features. *J. Appl. Meteorol*. **5**, 745–763.

Kimball, S. K., and M. S. Mulekar, 2004: A 15-Year Climatology of North Atlantic Tropical Cyclones, Part I: Size Parameters. *J. of Climate*, **17 (18)**, 3555–3575.

Malkin, W., and J.G. Galway, 1953: Tornadoes associated with hurricanes. *Mon. Wea. Rev.*, **81**, 299–303.

McCaul, E. W., Jr., 1991: Buoyancy and shear characteristics of hurricane--tornado environments. *Mon. Wea. Rev.*, **119**, 1954–1978.

McCaul, E.W., and M. L. Weisman, 1996: Simulations of shallow supercell storms in landfalling hurricane environments. *Mon. Wea. Rev.*, **124**, 408–429.

McCaul, E.W., and C. Cohen, 2002: The impact on simulated storm structure and intensity of variations in the mixed layer and moist layer depths. *Mon. Wea. Rev.*, **130**, 1722–1748.

McCaul, E.W., D. E. Buechler, S. J. Goodman, and M. Cammarata, 2004: Doppler radar and lightning network observations of a severe outbreak of tropical cyclone tornadoes. *Mon. Wea. Rev*., **132**, 1747–1763.

Moore, T.W., 2015: A statistical analysis of the association between tropical cyclone intensity change and tornado frequency. *Theoretical and Applied Climatology*, **1**–**11**.

Moore, T.W., and R.W. Dixon, 2011: Climatology of tornadoes associated with Gulf Coast-landfalling hurricanes. *Geogr. Rev.*, **101(3)**, 371–395.

National Weather Service, Central Pacific Hurricane Center, cited 2015: Hurricanes and Tornadoes. [Available online at http://www.prh.noaa.gov/cphc/pages/FAQ/Hurricanes_vs_tornadoes.php]

NOAA, 2008: Notes on Tropical Cyclone Tornadoes across Central NC. [Available online at http://www.erh.noaa.gov/rah/science/rah.tropical.cyclone.tornadoes.pdf]

Novlan, D.J. and W.M. Gray, 1974: Hurricane-spawned tornadoes *Mon. Wea. Rev.*, **102**, 476–488.

Onderlinde, M.J., and H.E. Fuelberg, 2014: A parameter for forecasting tornadoes associated with landfalling tropical cyclones. *Wea. Forecasting*, **29**, 1238–1255.

Orton, R., 1970: Tornadoes associated with Hurricane Beulah on September 19-23, 1967. *Mon. Wea. Rev.*, **98**, 541–547.

Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148-1164.

Schneider, Douglas, Scott Sharp, 2007: Radar Signatures of Tropical Cyclone Tornadoes in Central North Carolina. *Wea. Forecasting*, **22**, 278-286.

Schultz, L.A., and D.J. Cecil, 2009: Tropical cyclone tornadoes, 1950-2007. *Mon. Wea. Rev.*, **137**, 3471-3484.

Smith, J.S., 1965: The hurricane-tornado. *Mon Wea Rev*., **93**, 453-459.

Spratt, S. M., D. W. Sharp, P. Welsh, A. Sandrik, F. Alsheimer, and C. Paxton, 1997: A WSR-88D assessment of tropical cyclone outer rainband tornadoes. *Wea. Forecasting*, **12**, 479-501.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings with supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243-1261.

Verbout, S.M., Schulz, D.M., Leslie, L.M., Brooks, H.E., Karoly, D.J., and K.L. Elmore, 2007: Tornado outbreaks associated with landfalling hurricanes in the North Atlantic basin, 1954–2004. *Meteorol. Atmos. Phys*., **97**, 255-271.

Vescio, M. D., S.J. Weiss, and F. P. Ostby, 1996: Tornadoes associated with Tropical Storm Beryl. *Natl. Wea. Assoc. Dig.*, **21** (1), 2-10.

Weiss, S. J., 1985: On the Operational Forecasting of Tornadoes Associated with Tropical Cyclones. *Preprints, 14th Conference on Severe Local Storms*, Indianapolis, IN, 293-296.