ABSTRACT

GRAHAM, CHASE S. Verification of Convection-Allowing NWP in Southeastern U.S. High-Shear, Low-CAPE Environments. (Under the direction of Dr. Gary Lackmann).

Forecasting severe convection hazards such as tornadoes and damaging straight-line winds in environments with limited atmospheric instability and high amounts of vertical wind shear has proven to be an arduous task with life-threatening consequences in the event of an erroneous forecast. In the southeastern United States, these "High-Shear, Low-CAPE" (HSLC) environments tend to occur in the cool season and frequently occur at night, times when the general public's situational awareness of potential severe hazards is low. For these reasons, it is essential that forecasters be given as many effective tools as possible to try and successfully predict HSLC severe convection. This study examines the performance of one forecast metric which has shown success at predicting severe convection hazards in environments with large atmospheric instability, updraft helicity (UH), in HSLC environments.

In this study, High Resolution Ensemble Forecast (HREF) 0-3 km ensemble maximum UH guidance is verified against Storm Prediction Center (SPC) Local Storm Reports (LSRs) over 144 six-hour periods ("events") where HSLC severe convection was observed in the southeastern United States between 15 October and 15 April over the span of three cool seasons, 2017-18, 2018-19, and 2019-20. Neighborhood verification metrics such as the Fractions Skill Score (FSS) were employed to allow for some leniency for spatial discrepancies between the UH forecasts and the observations. Traditional verification methods including contingency table statistics were also used to determine more information about the nature of forecast errors. In addition to observing the overall performace of UH forecasts, this study also seeks to determine the relative performance between various UH thresholds in HSLC environments.

On average, all verification metrics suggest that UH guidance does not perform as well as in high-CAPE environments, with mean FSS and contingency table results placing forecast performace closer to no skill than a perfect forecast. The UH threshold of 35 m^2/s^2 was generally the most skillful, although 25 and 50 m^2/s^2 were fairly competitive. The highest UH threshold examined, 75 m^2/s^2, was clearly the least skillful on average, having an underforecasting bias. Although the lower UH thresholds were more skillful on average, they still possessed a high False Alarm Ratio (FAR), suggesting the opposite problem, an *overforecasting* bias. The intermediate UH thresholds were most successful primarily due to the fact that they displayed the

least forecast bias; however, some of the perceived strength in performance of the intermediate thresholds may be due to the way in which this study dealt with using LSRs. Weighting events by number of LSRs produced higher mean skill values for all thresholds, suggesting that UH forecast performance is better in situations with a large amount of LSR activity. Stratifying UH forecast performance by time of day did not produce any discernible trends; additionally, UH forecasts generally performed better when verifying against only wind reports than verifying against only tornado reports.

The findings of this study suggest that the benefits of using UH in HSLC environments may be limited for an average event, although UH forecasts did perform well for several of the events included in our database. Further, if forecasters are going to use UH as a tool for predicting HSLC severe convection hazards, the findings of this study suggest that they would be best served by using a threshold of between 25 and 50 m^2/s^2.

Verification of Convection-Allowing NWP in Southeastern U.S. High-Shear, Low-CAPE Environments

by
Chase S. Graham

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Marine, Earth and Atmospheric Sciences

Raleigh, North Carolina
2021

APPROVED BY:

Gary Lackmann Committee Chair	Sarah Larson
Matthew Parker	

DEDICATION

To my mom, dad, and brother, for always supporting me in pursuing my dreams.

BIOGRAPHY

Chase was born and raised in Dobson, North Carolina, and from an early age, had a passion for observing and forecasting the weather. All throughout elementary, middle, and high school, teachers and classmates would frequently ask for Chase's opinions on the upcoming forecast, especially in the wintertime when a chance for snow was on the horizon. This passion for meteorology led Chase to pursue a degree in Atmospheric Sciences from the University of North Carolina at Asheville. During his time at UNC-Asheville, Chase gained valuable experience from both classes as well as various research experiences, which varied from launching weather balloons during winter weather events to surveying tornado damage as part of the VORTEX-SE program to analyzing tropical cyclone precipitation patterns as part of the Ernest F. Hollings Scholarship program. Chase was also an active member of the American Meteorological Society student chapter at UNC-Asheville, participating in and organizing various social and community outreach activities. In 2018, Chase moved from his "home away from home" in the mountains to Raleigh to pursue graduate studies under Dr. Gary Lackmann at North Carolina State University. In his spare time, Chase enjoys outdoor activities including playing golf or disc golf with friends as well as musical activities such as singing or playing the piano. However, if winter weather or severe weather threaten, you can be certain that Chase will be glued to the radar with eager anticipation.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Gary Lackmann, for all the support and guidance that he has given me over the past three years. I am particularly thankful for his patience and understanding during the times when the research progress was slow, and for his persistent encouragement that he has given me to help me complete my studies. In addition, I would like to acknowledge my committee members – Drs. Sarah Larson and Matthew Parker – for their suggestions and contributions which will have improved this project.

This research was funded by grant NA17NWS4680002, as part of the NOAA CSTAR program. I would also like to acknowledge the contributions of the National Weather Service forecasters who participated in periodic CSTAR conference calls. The suggestions made by forecasters certainly helped to strengthen this thesis work and make it more useful in operations. I would like to thank Drs. Adam Clark and Brett Roberts from the Storm Prediction Center for helping me to access the HREF data, as well as Lindsay Blank from the Developmental Testbed Center for helping me with questions about the MET software package.

I would like to express my great appreciation for the assistance from fellow graduate students at NC State, specifically Trevor Campbell and Jacob Radford who helped me tremendously with codes and completing various research activities. I would also like to acknowledege the support which I received from other graduate students in the Marine, Earth, and Atmospheric Sciences department; this support helped to keep my spirits up during the difficult periods while at NC State.

Finally, I would like to thank my family for all the love and encouragement that you have given to me over the years. Without your support, none of the things that I have accomplished would have been possible.

TABLE OF CONTENTS

LIST OF TABLES vi

LIST OF FIGURES vii

Chapter 1 Introduction **1**
1.1 High-Shear, Low-CAPE (HSLC) Severe Convection 1
1.2 CAM Guidance in HSLC Environments 2
1.3 High Resolution Ensemble Forecast 3
1.4 Motivation and Thesis Outline 4

Chapter 2 Data and Methods **6**
2.1 HSLC Severe Convection Database 6
2.2 Forecasts 7
2.3 Observations 9
2.4 Verification Metrics 11
2.4.1 Fractions Skill Score 12
2.4.2 Contingency Table Statistics 14

Chapter 3 Results **25**
3.1 Fractions Skill Scores 25
3.1.1 Fractions Skill Scores for Selected Events 25
3.1.2 Aggregated Fractions Skill Scores 26
3.1.3 Variations in Fractions Skill Scores by UH Threshold 27
3.1.4 Variations in Fractions Skill Scores by LSR Type 29
3.1.5 Variations in Fractions Skill Scores by Time of Day 30
3.2 Contingency Table Statistics 31
3.2.1 Variations in Contingency Table Statistics by UH Threshold 33
3.2.2 Variations in Contingency Table Statistics by Time of Day 36

Chapter 4 Conclusions **59**
4.1 Summary of Findings 59
4.2 Comparison to Prior UH Verification Studies 61
4.3 Future Work 64

BIBLIOGRAPHY **66**

APPENDICES **72**

LIST OF TABLES

Table 1.1 Members of the HREF for versions 2 and 2.1. "HRW" refers to the High Resolution Window runs, while "NAM" refers to the North American Mesoscale Forecast System runs. "ARW" refers to the Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2008), "NMMB" refers to the Nonhydrostatic Multiscale Model on the B Grid (Janjić and Gall 2012), "NSSL" refers to the National Severe Storms Laboratory version of the WRF-ARW, and "HRRR" refers to the High Resolution Rapid Refresh member. 5

Table 2.1 Comparison between the average and median number of grid cells with UH exceeding a given threshold and the average and median number of LSRs across all 144 events. Because nearest neighbor methods are used for gridding LSRs, it can be assumed that 1 LSR = 1 grid cell with an LSR. 16

LIST OF FIGURES

Figure 2.1 Study Domain. The study domain is contained within the box demarcated by the two parallels, 29°N and 40°N and the two meridians, 94°W and 75°W. Areas over the Gulf of Mexico and Atlantic Ocean were not considered to be parts of the study domain and were not included in skill score calculations. 17

Figure 2.2 NEXRAD coverage by lowest beam height across the contiguous 48 states. Several regions within the study domain where the lowest beam height is above 6,000 feet are indicated by black ovals. Plot created by NOAA Radar Operations Center, https://www.roc.noaa.gov/WSR88D/. 18

Figure 2.3a Gridded LSR data (red dots) from 00-06 UTC 17 December 2019. This field was transferred onto the HREF grid using nearest-neighbor methods without any enhancement. 19

Figure 2.3b As in Figure 2.3a, but the gridded LSR data has been enhanced such that adjacent grid cells in a 7x7 area around the grid cell which contained the LSR are treated as if they also contained an LSR. 20

Figure 2.4 HREF Ensemble Maximum 0-3 km. Updraft Helicity (blue shading), 00-06 UTC, 17 December 2019. The disparity between the length and width of the UH swaths is denoted by the embedded text boxes. 21

Figure 2.5 Distributions of the 99th (left), 99.5th (center), and 99.9th (right) percentile HREF six-hour Ensemble Maximum Updraft Helicity values in regions where SBCAPE < 1000 J/kg. The distributions are taken from 109 six-hour observation periods spanning 51 cases. 22

Figure 2.6 Contingency Table with definitions for how forecasts and observations were separated into “yes” and “no” bins. Both the Updraft Helicity Threshold (“Threshold”) and Fractional Coverage Threshold (FCT) were manipulated for different tests. 23

Figure 2.7 Performance Diagram as described in Roebber (2009). Success Ratio (SR) is plotted along the abscissa, while Probability of Detection (POD) is plotted along the ordinate. Critical Success Index (CSI) values are determined by a combination of SR and POD as described by (11), and CSI contours are shown by the multicolored solid lines. Forecast bias as described by (12) is contoured by the dashed black lines. 24

Figure 3.1 HREF 6-hr. Ensemble Maximum Updraft Helicity (blue shading) and Enhanced LSRs (red squares) plotted for 00-06 UTC 17 December 2019. Updraft Helicity thresholds are denoted by differing shades of blue, with deeper blues representing higher UH values. LSRs are enhanced using the 7x7 enhancement discussed in Section 2, which is what is used in verification. 39

Figure 3.2 As in Figure 3.1, but for 06-12 UTC 6 February 2020. 40

Figure 3.3 As in Figure 3.1, but for 00-06 UTC 23 January 2018. 41

Figure 3.4 Number of events for which a specified threshold had the highest Fractions Skill Score (FSS) (Total Number of Events = 144). The "None/Multiple Thresholds" category includes events where either multiple thresholds had equally high FSS values or all thresholds had an FSS value of 0. 42

Figure 3.5 Number of events for which a specified threshold had the lowest Fractions Skill Score (FSS) (Total Number of Events = 144). The "None/Multiple Thresholds" category includes events where either multiple thresholds had equally low FSS values or all thresholds had an FSS value of 0. 43

Figure 3.6 Mean FSS values by UH threshold over all events (n = 144), verifying against tornado and damaging wind storm reports. UH threshold increases from left to right, beginning with 25 m^2/s^2 (red), then 35 m^2/s^2 (blue), 50 m^2/s^2 (green), and finally 75 m^2/s^2 (yellow). The mean values for this plot were calculated weighting each event equally. 44

Figure 3.7 As in Figure 3.6, but mean FSS values were calculated by weighting each event depending on the total number of tornado and damaging wind reports that event had. See Section 3.1.2 and Equation 3.2 for more information on the weighting method. 45

Figure 3.8 Mean FSS values by observation dataset over all events (n = 144). Verification against "All LSRs" (no outline) includes both tornado and damaging wind reports, with "Only Wind LSRs" (black outline) representing verification solely against damaging wind storm reports and "Only Tornado LSRs" (cyan outline) representing verification solely against tornado reports. The mean values for this plot were calculated weighting each event equally. 46

Figure 3.9 As in Figure 3.8, but mean FSS values were calculated by weighting each event depending on the total number of tornado and damaging wind reports that event had. See Section 3.1.2 and Equation 3.2 for more information on the weighting method. 47

Figure 3.10 Mean FSS values, grouped by the six-hour window in which the event took place. The color scheme denoting which UH threshold is used is consistent with Figures 6-9. The leftmost grouping of bars represents the mean FSS for events which took place in the 12-18 UTC time frame, with groupings proceeding through the diurnal cycle from left to right. The mean values for this plot were calculated weighting each event equally. 48

Figure 3.11 As in Figure 3.10, but mean FSS values were calculated by weighting each event depending on the total number of tornado and damaging wind reports

that event had. See Section 3.1.2 and Equation 3.2 for more information on the weighting method. 49

Figure 3.12 Performance Diagram using a minimum Fractional Coverage Threshold. Probability of Detection (POD) and Success Ratio (SR) ordered pairs are plotted for all events (*n* = 144) at the specified UH thresholds. Individual events are represented by the small dots, while the equally-weighted (black outline) and storm report-weighted (grey outline) mean POD and SR values are denoted by the large dots. The color scheme is based on UH threshold and is the same as in Figures 3.3-3.8. Lines of equal Critical Success Index (CSI) are indicated by the multicolored contours, starting from CSI = 0.1 and labeled at an interval of 0.1 up to CSI = 0.9. 50

Figure 3.13 As in Figure 3.12, but for a 1% Fractional Coverage Threshold. 51

Figure 3.14 As in Figure 3.12, but for a 2% Fractional Coverage Threshold. 52

Figure 3.15 As in Figure 3.12, but for a 5% Fractional Coverage Threshold. 53

Figure 3.16 As in Figure 3.12, but for a 10% Fractional Coverage Threshold. 54

Figure 3.17 Performance Diagram using a minimum Fractional Coverage Threshold, highlighting variations in POD and SR values for events at different portions of the diurnal cycle. POD and SR ordered pairs are plotted for all events (n = 144) using the same UH threshold; in this plot, UH ≥ 25 m^2/s^2. Individual events are represented by the small dots, while the equally-weighted (black outline) and storm report-weighted (grey outline) mean POD and SR values are denoted by the large dots. The color scheme of the dots is based on the six-hour window which an event falls into, including 12-18 UTC (maroon, n = 30), 18-00 UTC (orange, n = 47), 00-06 UTC (chartreuse, n = 39), 06-12 UTC (cyan, n = 28). CSI contours are plotted as in Figure 3.9. 55

Figure 3.18 As in Figure 3.17, but for the threshold UH ≥ 35 m^2/s^2 56

Figure 3.19 As in Figure 3.17, but for the threshold UH ≥ 50 m^2/s^2 57

Figure 3.20 As in Figure 3.17, but for the threshold UH ≥ 75 m^2/s^2 58

CHAPTER 1

Introduction

1.1 High-Shear, Low-CAPE (HSLC) Severe Convection

Severe convection produces an array of hazards which threaten life and property, from unpredictable microscale phenomena such as individual lightning strikes to more predictable mesoscale phenomena such as areal flooding. The environmental paradigm that has been associated with the bulk of the previous severe convection research includes large amounts of vertical wind shear as well as high atmospheric instability, and typically takes place in the spring and early summer across the Great Plains region of the United States. However, severe convection and its associated hazards can also occur in environments with limited atmospheric instability, typically referred to as "High-Shear, Low-CAPE" (HSLC) environments. These environments, while existing at all times of year and in nearly all regions of the continental United States, have a tendency to occur and produce severe hazards during the cool season in the Southeastern, Lower Mississippi Valley, and Ohio Valley regions of the United States (Schneider et al. 2006; Sherburn and Parker 2014; Sherburn et al. 2016). Additionally, HSLC severe convection frequently occurs during the overnight hours (Sherburn and Parker 2014; Sherburn et al. 2016). From a societal standpoint, this is problematic for two reasons: first, during the cool season (and especially at night!), situational awareness of severe weather threats is lower than during the months typically associated with high-CAPE severe convection, and second, the regions where HSLC severe convection tends to occur have a preponderance of mobile home housing in comparison to the rest of the United States, which places a greater number of people at risk of serious injury or death when severe convection hazards occur (Ashley et al. 2008; Strader and Ashley 2018).

Considering the unfavorable societal factors that are associated with HSLC severe convection, forecasters are under greater pressure to make accurate forecasts and warnings during these events. Unfortunately, forecasting and warning HSLC severe convection events incurs unique challenges, as HSLC severe hazards are often quite ephemeral in nature (Davis and Parker 2014), and are frequently imbedded within larger structures such as quasi-linear convective systems (QLCS) (Thompson et al. 2012). Additionally, as a byproduct of the limited

amounts of instability, HSLC severe convection exhibits a smaller horizontal scale and is less vertically extensive compared to severe convection in high-CAPE regimes, which makes potential severe threats such as tornadoes or damaging straight-line winds more difficult to observe with the WSR-88D network (Davis and Parker 2014). As a result of these factors, forecast and warning skill of severe hazards in HSLC environments lags behind that of high-CAPE environments, with forecasts and warnings having lower probability of detection (POD) and higher false alarm ratios (FAR) (Dean and Schneider 2008; Anderson-Frey et al. 2016).

1.2 CAM Guidance in HSLC Environments

In recent years, several efforts have been made to provide forecasters with additional tools with the ultimate goal of improving forecast skill in HSLC environments. Studies have attempted to analyze meso- and synoptic-scale differences in various environmental parameters between HSLC environments that did and did not produce hazards associated with severe convection (King et al. 2016; Sherburn et al. 2016). In addition, efforts have been made to create environmental parameters which use various model output fields to skillfully forecast HSLC severe convection events (Sherburn and Parker 2014; Sherburn et al. 2016). While these studies have undoubtedly provided forecasters with additional information and resources that can be used in prediction of HSLC severe convection events, the advent of convection-allowing model (CAM) guidance provides the opportunity to give forecasters another set of resources which can be used in the prediction of severe hazards in HSLC environments. Further, CAM guidance can provide information that bridges the gap from short-term (1-3 day lead time) environmental-scale prediction to nowcasting tools such as satellite and radar, hopefully reducing the amount of errors which is prevalent in HSLC severe watches. However, it is essential that any new CAM guidance fields which are to be used to predict areas of convection and associated hazards be tested to ensure that they are skillful.

In spite of recent improvements in horizontal and vertical resolution provided by CAMs, even the most high-resolution operationally-available guidance is too coarse to fully resolve individual severe hazards such as tornadoes or swaths of damaging straight-line winds (Bryan et al. 2003). However, severe convection “proxies” have been developed which predict convective structures which are typically associated with severe hazards, such as mesocyclones. Kain et al. (2008) first suggested the use of Updraft Helicity, a severe proxy which combines vertical

vorticity and updraft strength to represent the occurrence of rotating mesocyclones. Several verification studies have examined the skill of these forecasts in various temporal and spatial setups, generally finding updraft helicity to have good skill as a proxy predictor of severe convection hazards such as tornadoes, damaging straight-line winds, and hail (e.g., Sobash et al. 2011; Schwartz et al. 2015; Sobash et al. 2016; Dawson et al. 2017; Sobash and Kain 2017). While updraft helicity has been the preeminent severe convection proxy, other forecast fields such as maximum vertical vorticity and reflectivity have also been examined as a potential proxy for severe convection hazards (Yussouf et al. 2013; Yussouf et al. 2015; Skinner et al. 2016; Roberts et al. 2020), although the focus of this study will be on verification of updraft helicity.

Because of the spatially-confined nature of the forecast problem that is associated with prediction of severe convection, many studies have resorted to using neighborhood verification approaches, which allow for modest amounts of spatial error between severe convection proxy forecasts and observations. Schwartz et al. (2010) first proposed the use of the Fractions Skill Score (Roberts (2005); Roberts et al. 2008) as a neighborhood verification metric of severe convection forecasts, and many subsequent severe convection verification studies (including most of those mentioned in the preceding paragraph) have also employed this metric. One challenge that using neighborhood verification methods presents is that forecast skill varies based on the size of the neighborhood used. Therefore, it is important that neighborhood size aligns with the amount of error that would be tolerated for a forecast to still be deemed successful. Another verification approach which is frequently used in severe convection forecast verification is the "practically perfect" method, which takes observations of severe convection (typically storm reports) and uses a Gaussian filter to smooth the storm reports into a probabilistic "hindcast" which would represent the forecast that would be made if a forecaster had perfect knowledge of how a severe convection event would transpire beforehand (Hitchens et al. 2013). Initially, this method was used to verify Storm Prediction Center probabilistic outlooks for tornadoes, hail, and damaging winds, although recent studies have examined using practically perfect verification with NWP forecasts (e.g., Gensini and Tippett 2019).

1.3 High Resolution Ensemble Forecast

The creation of the first operationally active CAM ensemble began as an amalgamation of seven deterministic CAMs known as the Storm-Scale Ensemble of Opportunity (SSEO; Jirak

et al. 2012). Instead of configuring an ensemble by varying the initial and/or boundary conditions of a single model, the SSEO incorporated the output a several different CAMs with differing dynamical cores, grid resolutions, and microphysical parameterizations. Testing of the performance of the SSEO was done yearly during the Spring Forecasting Experiment (e.g. Gallo et al. 2017), with adjustments in the ensemble's configuration ultimately leading to the creation and operational implementation of the High Resolution Ensemble Forecast version 2 in November 2017 (HREFv2; Roberts et al. 2019).

The HREFv2 consisted of eight different ensemble members based on four different model configurations: the HRW NSSL and HRW ARW members and their 12 hour time-lagged runs, which use the Advanced Research version of the Weather Research and Forecasting (WRF-ARW; Skamarock et al. 2008) model core, and the HRW NMMB and NAM Nest members and their 12 hour time-lagged runs, which use the Nonhydrostatic Multiscale Model on the B Grid (NMMB; Janjić and Gall 2012) model core. The HREF configuration was modified to its present operational form (HREFv2.1) in April 2019 with the addition of the HRRR and its 6-hour time-lagged run, which use the WRF-ARW model core. Since its inception, the HREF has had two runs per day, one at 00 UTC and one at 12 UTC, and HREF forecasts extend out 48 hours (36 hours for the HRRR runs). HREF data is available to forecasters as well as the public via the Storm Prediction Center website (https://www.spc.noaa.gov/exper/href/). While this study will focus on updraft helicity, other HREF severe convection proxies such as maximum reflectivity, maximum updraft strength, and maximum 10-meter wind speed are also available to forecasters.

1.4 Motivation and Thesis Outline

Personal discussions and informal survey have revealed that the vast majority of National Weather Service forecasters use some form of CAM guidance in their forecast process. However, without verification studies, forecasters are using a tool of unknown skill. While there have been several studies to assess the skill of severe convection proxies such as updraft helicity, none of the antecedent literature has exclusively focused on updraft helicity performance in HSLC environments. This study will make a first attempt at doing that, aiming to provide forecasters with initial guidance related to the usefulness and applicability of various updraft helicity thresholds during HSLC events in the southeastern United States.

In Section 2, the data and methods used for this study will be discussed, including the process for selecting HSLC severe convection events, the forecast and observation data used, and the verification metrics employed. Section 3 will discuss the results of the study, assessing overall performance of updraft helicity forecasts as well relative performance between different updraft helicity thresholds. Finally, Section 4 will summarize the key findings of section 3 and make recommendations for forecasters, put the findings of this study in the context of past verification studies, and discuss potential future work and improvements which may be made to this study.

Table 1.1. Members of the HREF for versions 2 and 2.1. "HRW" refers to the High Resolution Window runs, while "NAM" refers to the North American Mesoscale Forecast System runs. "ARW" refers to the Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2008), "NMMB" refers to the Nonhydrostatic Multiscale Model on the B Grid (Janjić and Gall 2012), "NSSL" refers to the National Severe Storms Laboratory version of the WRF-ARW, and "HRRR" refers to the High Resolution Rapid Refresh member.

Member	**Included in HREF Hours**	**Included in v2/v2.1?**	**Has 0-3 km UH?**
HRW ARW	0 – 48	Yes/Yes	Yes
HRW ARW (-12 h)	0 – 36	Yes/Yes	Yes
HRW NMMB	0 – 48	Yes/Yes	Yes
HRW NMMB (-12 h)	0 – 36	Yes/Yes	Yes
HRW NSSL	0 – 48	Yes/Yes	Yes
HRW NSSL (-12 h)	0 – 36	Yes/Yes	Yes
NAM Nest	0 – 48	Yes/Yes	No
NAM Nest (-12 h)	0 – 36	Yes/Yes	No
HRRR	0 – 36	No/Yes	Yes
HRRR (-6 h)	0 – 30	No/Yes	Yes

CHAPTER 2

Data and Methods

2.1 HSLC Severe Convection Database

One of the first steps which was undertaken was the configuration of a database of cases where HSLC severe convection occurred. The temporal and spatial extent of the study and, by extension, the areas and window of time over which potential cases were gathered was limited to the southeastern United States during the cool season, which is defined in this study as between 15 October and 15 April. For the purposes of this study, the "southeastern United States" includes an area bounded by 29°N on the south, 40°N on the north, 94°W on the west, and 75°W on the east (Figure 2.1). All areas over large bodies of water such as the Atlantic Ocean or the Gulf of Mexico were not included in the study domain. In addition to confining the study to the cool season, the study was also confined to three seasons: 2017-18, 2018-19, and 2019-20. This was purely a practical constraint: the HREF only became operational on 1 November 2017 (this means that cases from the first two weeks of the 2017-18 cool season cannot be used).

The first step in the process of creating a HSLC severe convection database involved parsing through archived daily files of Storm Prediction Center (SPC) Local Storm Reports (LSRs). If there was at least one report of damaging straight-line winds or tornadoes within an SPC day (12 UTC - 12 UTC) in the study domain, that day would be considered as a potential event. Next, archived NEXRAD data were used to confirm the accuracy of the LSRs. Despite the SPC's best efforts to filter the LSR data, there were occasionally reports which were not associated with any substantial reflectivity returns and were likely erroneous. In most cases, it appeared that these "false" LSRs were attributed to a location which may have experienced severe weather, but on the day before or after the day that it was reported. These reports were considered invalid, and as a result, some potential events were removed.

After these quality control measures were conducted, the potential events were stratified by whether or not they occurred in environments that contained large amounts of instability. Archived hourly SPC Mesoanalysis most unstable parcel CAPE (MUCAPE) fields were subjectively analyzed to determine whether the environment for a potential case could be considered high-CAPE or low-CAPE. In this study, a "low-CAPE" environment is one that has

MUCAPE of less than or equal to 1000 J/kg, in order to maintain consistency with prior HSLC definitions (e.g. Sherburn and Parker 2014; King et al. 2017). While previous studies (e.g., Schneider et al. 2008, Sherburn and Parker 2014) included a wind shear criterion in their definition of HSLC, in this study, we only use an instability criterion, with the assumption that low instability cool season environments need to have large amounts of dynamic forcing in order to produce severe convection. The three different categories that potential events were broken into were as follows: events which were HSLC (MUCAPE ≤ 1000 J/kg) during no portion of the event, events which were HSLC either in part of the domain or for part of the duration of the event, and events which were HSLC during the entirety of the event. Over the three cool seasons, October through April, 2017-2020, there were a total of 150 potential events; in other words, there were 150 SPC days that contained at least one valid storm report, regardless of whether the event satisfied the instability criterion. 57 potential events were HSLC during no portion of the event, and were removed from consideration. Of the remaining 93 events, 31 events were HSLC during the entirety of the event, while 62 events were HSLC for at least some portion of the event. These events which were either wholly or partially HSLC were included in the HSLC severe convection events database and were the ones ultimately used in the verification study. It should be noted that while there were initially 93 different HSLC cases, there were some occurrences where consecutive days with HSLC severe convection were actually grouped together and considered as one event. Additionally, data for some events was not readily available, which resulted in fewer than 93 separate events being examined by the time the verification was conducted.

2.2 Forecasts

Despite improvements in the horizontal resolution of operational high-resolution forecast models, model data are still too coarse to fully resolve convective-scale processes and hazards such as individual mesocyclones or tornadoes (Bryan et al. 2003). However, severe convection proxies have been developed in order to try and predict where severe hazards may occur using grid-scale variables. The main severe convection proxy which was used in this study is Updraft Helicity (UH; Kain et al. 2008), which attempts to predict locations of mesocyclones or rotating updrafts using the product of vertical velocity (w) and the vertical component of vorticity (ζ) over a given layer bounded at an upper (Z_U) and lower (Z_L) height level:

$$UH = \int_{Z_L}^{Z_U} w\zeta \, dz \qquad (2.1)$$

Since the strength of convective phenomena like those represented by UH ebb and flow over time scales much smaller than the difference between output times of even high-resolution models, it is important to be able to examine the evolution of the UH field between model output times. Following the recommendations of Kain et al. (2010), instead of using the instantaneous values of UH at the model output times, this study uses hourly maximum UH, which produces swaths of UH more similar in nature to actual storm tracks. Further, in order to allow for some leniency in time as well as more closely resemble the products used operationally, UH forecasts were aggregated over six-hour intervals (Figure 2.4). After exhausting the database of HSLC severe convection events and filtering to ensure that false events with erroneous storm reports were not included, there were a total of 144 six-hour intervals that were processed for verification.

The choice of levels for the upper and lower boundaries of the layer over which UH was integrated was not trivial for this study. Practically all previous severe convection verification studies which used UH as a severe convection proxy used UH calculated over a layer between 2 km and 5 km AGL (i.e., "2-5 km UH") (e.g., Sobash et al. 2011; Sobash et al. 2016; Dawson et al. 2017). In contrast, since this study is specifically looking at severe convection verification in low-CAPE environments, and many HSLC vortices are contained below the 2-5 km layer (Parker and Davis 2014; Wade 2020), it was important that, if possible, UH calculated over more shallow layers be used as the predictor of severe convection. As a result of these considerations, for this study, 0-3 km UH was used as the proxy for forecasted severe convection. All members of the HREF v2 and v2.1 contained 0-3 km UH as a forecast variable with the exception of the NAM Nest members. Other than the exclusion of the NAM Nest members, the configuration of the HREF used in this study followed that which was used operationally. HRW-ARW, HRW-NMMB, and HRW-NSSL members (as well as their 12-hour time-delayed runs) were included in the ensemble for all events between 1 November 2017 until 31 March 2019, while the HRRR member (and its 6-hour time-delayed run) was added to all HREF forecasts for all events within study window on or after 1 April 2019. As was discussed in Chapter 1, for each event, there was

a choice between using the 00 UTC and the 12 UTC HREF run. While no explicit guidelines were set to determine which run to use, as a general rule, the HREF run used in verification was the one closest to the onset of an event without being so close that "nowcasting" techniques (i.e., radar/satellite analysis) would be used. Instead of evaluating the performance of each of the members separately as if they were deterministic forecasts and averaging the results, ensemble fields were created using the ensemble-stat tool in the Model Evaluation Tools (MET; Jensen et al. 2020) 9.0.1 software package. There are difficulties in using certain HREF ensemble fields such as Ensemble Probabilities and Ensemble Mean values because each of the HREF member solutions are not equally likely, and additional statistical processing and bias correction would have to take place before those fields could be verified. Since this is a first attempt at verification of the HREF, for the sake of simplicity, Ensemble Maximum UH fields were chosen as the forecast dataset for this study.

2.3 Observations

Multiple datasets were considered for use as a representation of "observed severe convection", including rotation track data from the Multi-Radar/Multi-Sensor System (MRMS). While there were benefits to using the MRMS rotation track data for the observational dataset (continuous field, similar in scale to UH), there are also notable gaps in the radar coverage below 6,000 feet within the study domain (Figure 2.2). Additionally, past studies have shown that there are difficulties in differentiating between tornadic and non-tornadic radar signatures in HSLC environments, especially outside of 60 km from a radar site (Davis and Parker 2014). Instead, archived SPC LSR data were obtained (www.spc.noaa.gov/archive) for use as the observational dataset. While previous studies examining model performance in High-CAPE environments (Sobash et al. 2011, Dawson et al. 2017) used all LSR types, we do not expect hail greater than 19 mm (¾ in.) to be a severe hazard frequently associated with low-CAPE convection; therefore, the only LSR types used in the observation dataset were tornado reports and reports of damaging wind gusts in excess of 26 m s^{-1} (58 mi. (hr^{-1})). This assumption about a lack of hail reports was generally true for our list of cases: the total amount of hail reports (363) was an order of magnitude less than the amount of damaging wind and tornado reports (3148). As a result, excluding these hail reports is unlikely to have a great impact on the overall skill results. To allow for small temporal discrepancies between the HREF UH forecast data and the LSRs, the

LSRs are also aggregated into six-hourly windows. One documented limitation of LSRs is underreporting of severe hazards in sparsely populated regions (e.g., Potvin et al. 2019). While this may be of concern in some regions within the domain (for example, areas in and around the Appalachian Mountains), overall, LSRs appear to be the most reliable dataset for observing severe hazards, as well as a dataset which has been implemented in similar antecedent studies (Sobash et al. 2011; Dawson et al. 2017).

One challenge of using LSRs in verification is that they must be transferred onto a grid in order to be directly compared to the gridded forecast (UH) data. For this study, LSR data were mapped onto the HREF grid using nearest-neighbor methods. In essence, a binary observation field was created, where grid cells which contained an LSR were given a value of 1 and grid cells that did not contain an LSR were given a value of 0. When you compare an example forecast (UH) and observation (gridded LSRs) field from 00-06 UTC 17 December 2019 (Figures 2.3a and 2.4), a limitation in this methodology is revealed. It is clear that there is a considerable scale difference in the size of the UH swaths and the gridded LSRs, with UH swaths overspreading significantly larger areas than the gridded LSRs. Table 2.1 confirms this, revealing large differences between the average and median amount of grid cells which exceed the various UH thresholds and the average and median amount of LSRs. As a result, despite a forecast where the majority of LSRs are roughly colocated with the UH swaths, it appears that the HREF UH field is overforecasting severe convection. In order to overcome the difference in the sizes of the UH and gridded LSR fields, the sizes of the gridded LSR field were artificially enhanced, something that has been done in prior studies dealing with verification of UH with LSRs (Dawson et al. 2017). With the enhanced LSRs, an area of seven grid cells by seven grid cells (21-km by 21-km) was given a value of 1 in the binary observation field for each LSR, with the grid cell nearest to the location of the LSR being the center grid cell. Because the UH swaths are more or less spatially expansive based on the particular choice of threshold (with low UH thresholds producing more spatially expansive fields and vice versa), different amounts of LSR enhancement will be more or less favorable for different thresholds. However, the choice of using a seven-by-seven area (an enhancement factor of 49) makes the scale of the LSRs somewhere in between the low UH thresholds and the high UH thresholds. While this is not an ideal setup, this accomplishes the task of making the scale of the forecasts and observations more equal without necessarily favoring the higher or lower UH thresholds. Figures 2.3a and 2.3b

show the difference between the LSR footprint for gridded LSRs that were unenhanced versus those that were enhanced. The enhanced gridded LSR field was used for verification of all events.

2.4 Verification Metrics

For this study, it was important to employ verification metrics and methodologies that were appropriate given the nature of the phenomenon being investigated. Since UH is a proxy which attempts to represent rotating mesocyclones, its spatial scale is typically only a few kilometers in the cross-storm direction, while ranging from a few kilometers to over 100 km in the direction of storm propagation (Figure 2.4). Additionally, after consulting with NWS meteorologists, it is evident that forecasters tend to use UH to determine general areas where severe convection may occur instead of interpreting the forecast as an exact representation of where severe convection will occur. As a result, it is important to allow the model some leniency both in space and time. As was mentioned previously, temporal leniency is granted through the aggregation of both the forecast and observation data into six-hour periods.

We allow for spatial leniency by using neighborhood verification metrics; instead of doing a direct comparison of the UH field to the field of gridded LSRs at each grid box, we compare the forecast and observation field over an area comprising multiple grid boxes. One challenge that this presents is that the forecast field (UH) is continuous, while the observation field (gridded LSRs) is binary. In order to conduct verification, we had to convert the continuous forecast data to a field of binary probabilities (BPs). This was done by thresholding the UH field at different values (UH_{thresh}) such that

$$BP_i = \begin{cases} 1, & UH_i \geq UH_{thresh} \\ 0, & UH_i < UH_{thresh} \end{cases} \qquad (2.2)$$

at the *i*th grid point. The threshold values selected for verification were determined by calculating the 99th, 99.5th, and 99.9th percentile value of UH in a manner with some similarities to that outlined in Dawson et al. (2017). The percentile values used are high because of the limited nature of UH swaths, which rarely occupy more than 1-2% of the domain. Essentially, a UH "climatology" was created by using data from a selection of events comprising

roughly three-quarters of the total database (n=109 6-hr. obs. periods, 51 events). For each six-hour observation period, the 99th, 99.5th, and 99.9th percentile values of UH were calculated within the study domain in regions where the ensemble mean surface-based CAPE was less than 1000 J/kg, in order to try to ensure that UH was generated in a low-CAPE regime. After aggregating the percentile values of UH over 109 observation periods, the distributions of the 99th, 99.5th, and 99.9th UH value were as shown in Figure 2.5. Ultimately, we decided to set our threshold values for verification at levels which roughly corresponded to the median value of each of these distributions. The median 99th percentile UH value (27.12 m^2/s^2) was rounded down to 25 m^2/s^2, and served as the lowest UH threshold, the median 99.5th percentile UH value (32.98 m^2/s^2) was rounded up to 35 m^2/s^2 and served as an intermediate UH threshold, and the median 99.9th percentile UH value (46.03 m^2/s^2) was rounded up to 50 m^2/s^2 and served as a high UH threshold. Since 75 m^2/s^2 is a threshold value used in UH products (albeit 2-5 km UH) on the SPC HREF webpage, we also decided to implement this threshold value into our verification process.

After the continuous UH field was converted to a binary field, we were then able to directly compare the HREF forecast (thresholded UH) to the observations (gridded LSRs) over a given area (a.k.a. the "neighborhood"). Both the neighborhood forecast probability as well as the neighborhood observation probability at a given grid box (NP_i) were calculated using

$$NP_i = \frac{1}{N_b} \sum_{m=1}^{N_b} BP_m \qquad (2.3)$$

where N_b is the total number of grid boxes within a neighborhood and BP_m is the binary probability of a forecast of $UH \geq UH_{thresh}$ or the binary probability of an observed gridded LSR at any given *m*th point within a neighborhood. Past work has shown that neighborhood size is not trivial and that forecast skill improves significantly as neighborhood size increases (Roberts and Lean 2008; Dawson et al. 2017). In order to be as consistent as possible with past studies which used either 80-km grids or 80-km x 80-km neighborhoods (e.g., Sobash et al. 2016; Roberts et al. 2020), our neighborhood size is set at 75-km x 75-km for all neighborhood verification calculations.

2.4.1 Fractions Skill Score

The skill score predominantly used in this study is the Fractions Skill Score (FSS; Roberts 2005; Roberts and Lean 2008). The FSS is a measure of the skill of a forecast relative to a forecast with no skill

$$FSS = 1 - \frac{FBS}{FBS_{worst}} \qquad (2.4)$$

As it relates to this study, FBS is the Fractions Brier Score of the HREF forecast and FBS_{worst} is the Fractions Brier Score of a hypothetical forecast with no skill.

$$FBS = \frac{1}{N_b}\sum_{m=1}^{N_b}(P(f)_m - P(o)_m)^2 \qquad (2.5)$$

$$FBS_{worst} = \frac{1}{N_b}\left(\sum_{m=1}^{N_b}P(f)_m^2 + \sum_{m=1}^{N_b}P(o)_m^2\right) \qquad (2.6)$$

$$FSS = 1 - \frac{\frac{1}{N_b}\sum_{m=1}^{N_b}\left(P(f)_m - P(o)_m\right)^2}{\frac{1}{N_b}\left(\sum_{m=1}^{N_b}P(f)_m^2 + \sum_{m=1}^{N_b}P(o)_m^2\right)} \qquad (2.7)$$

Considering that both the forecast and observation dataset are binary probability fields of values that are either 0 or 1, the sum of the squared forecast probability $P(f)^2$ and the squared observation probability $P(o)^2$ will be 1 at every grid box. Therefore, generalizing across the entirety of a neighborhood, (5) can be simplified to

$$FSS = 1 - \frac{\frac{1}{N_b}\sum_{m=1}^{N_b}\left(P(f)_m - P(o)_m\right)^2}{\frac{1}{N_b}N_b} = 1 - \frac{\sum_{m=1}^{N_b}\left(P(f)_m - P(o)_m\right)^2}{N_b} = 1 - MSE \qquad (2.8)$$

In this special case, the FSS is essentially equal to the difference of one and the mean squared error (MSE) over the entirety of the neighborhood. Since the maximum possible mean squared error when the forecasts and observations are binary probabilities of either 0 or 1 would be 1, the worst possible FSS would be 0 and is considered to be a forecast with no skill, while a perfect forecast has an FSS of 1. As was mentioned previously, FSS is a relative measure of skill

compared to a worst-case forecast and therefore has no intrinsic cutoff value where a forecast can be considered skillful. However, for the purposes of this study, we use FSS to try and determine which threshold value of UH produces the most skillful predictions of severe convection. The grid-stat verification tool in MET was used in the calculation of FSS values for all cases.

2.4.2 Contingency Table Statistics

While the Fractions Skill Score provides substantial information about the quality of forecasts using neighborhood verification methods, there are other metrics which can provide information about forecast performance that the FSS cannot. The 2 x 2 contingency table can be used to calculate additional statistics given two discrete outcomes for both the forecast as well as the observation field, typically represented as a "yes" or "no" outcome. Considering the nature of this study, our previously-employed UH thresholds (UH_{thresh}) were used to separate what qualifies as a "yes" or "no" forecast, with any UH exceeding the threshold being considered a "yes forecast" and vice versa. Handling of the observations was a simpler matter, with any occurrence of an LSR (enhanced or unenhanced) being considered a "yes observation" and vice versa. It was still important to allow for reasonable spatial discrepancies between forecasts and observations when calculating contingency table statistics, so the neighborhood methods employed in the calculation of FSS statistics were also used in the calculation of contingency table statistics. This was done by setting a fractional coverage threshold which had to be exceeded by both the forecasts and the observations for the neighborhood as a whole to be considered a "yes forecast" or "yes observation". For example, if a 1% fractional coverage threshold was set, for the application of this study, at least 1% of the grid cells within the 75-km x 75-km neighborhood would have to have UH exceeding the set threshold in order for the grid cell at the center of the neighborhood to be considered a "yes forecast" (Figure 2.6).

While different configurations of contingency table metrics can be used for a wide variety of applications, the combination of two metrics - probability of detection (POD) and false alarm ratio (FAR) - can be combined to provide an assessment of both the skill and bias of forecasts.

$$POD = \frac{a}{a+c} \qquad (2.9)$$

$$FAR = \frac{b}{a+b} \qquad (2.10)$$

Using a performance diagram (Roebber 2009), one plot can show not only the POD and FAR (more accurately, the Success Ratio (SR)), but also the Critical Success Index (CSI; also known as the "Threat Score") as well as the bias..

$$SR = 1 - FAR \qquad (2.11)$$

$$CSI = \frac{a}{a+b+c} = \frac{1}{\frac{1}{SR} + \frac{1}{POD} - 1} \qquad (2.12)$$

$$bias = \frac{a+b}{a+c} = \frac{POD}{SR} \qquad (2.13)$$

Figure 2.7 provides an example of how POD, FAR/SR, CSI, and bias can all be presented together graphically. SR is plotted along the abscissa, while POD is plotted along the ordinate. Both quantities range from 0 at the origin to 1 at the upper limit of their respective axes. A combination of the SR and POD quantities given by (2.12) is used to determine the CSI value. CSI values are near 0, suggesting a poor forecast, in the bottom left corner of a performance diagram as well as along the two axes. CSI values increase toward 1 as you move toward the upper right corner of the performance diagram, suggesting a more skillful forecast. Finally, the forecast bias is the ratio of POD to SR, as given by (2.13). Therefore, where SR is high and POD is low, such as along the abscissa, the forecast bias will be less than 1, suggesting an underforecast. In contrast, where POD is high and SR is low, such as along the ordinate, the forecast bias will be greater than 1, suggesting that an overforecasting bias is present.

For the purposes of this study, the performance diagrams will supplement the verification results provided by the FSS statistics, providing additional information about the nature of errors in HREF UH forecasts on a case-to-case basis that the single-valued FSS cannot.

Table 2.1. Comparison between the average and median number of grid cells with UH exceeding a given threshold and the average and median number of LSRs across all 144 events. Because nearest neighbor methods are used for gridding LSRs, it can be assumed that 1 LSR = 1 grid cell with an LSR.

	Average Grid Cells above UH Threshold/Number of LSRs	**Median Grid Cells above UH Threshold/Number of LSRs**
LSRs	22	7.5
UH ≥ 25 m^2/s^2	4198	2235
UH ≥ 35 m^2/s^2	1954	649.5
UH ≥ 50 m^2/s^2	686	138
UH ≥ 75 m^2/s^2	145	13.5

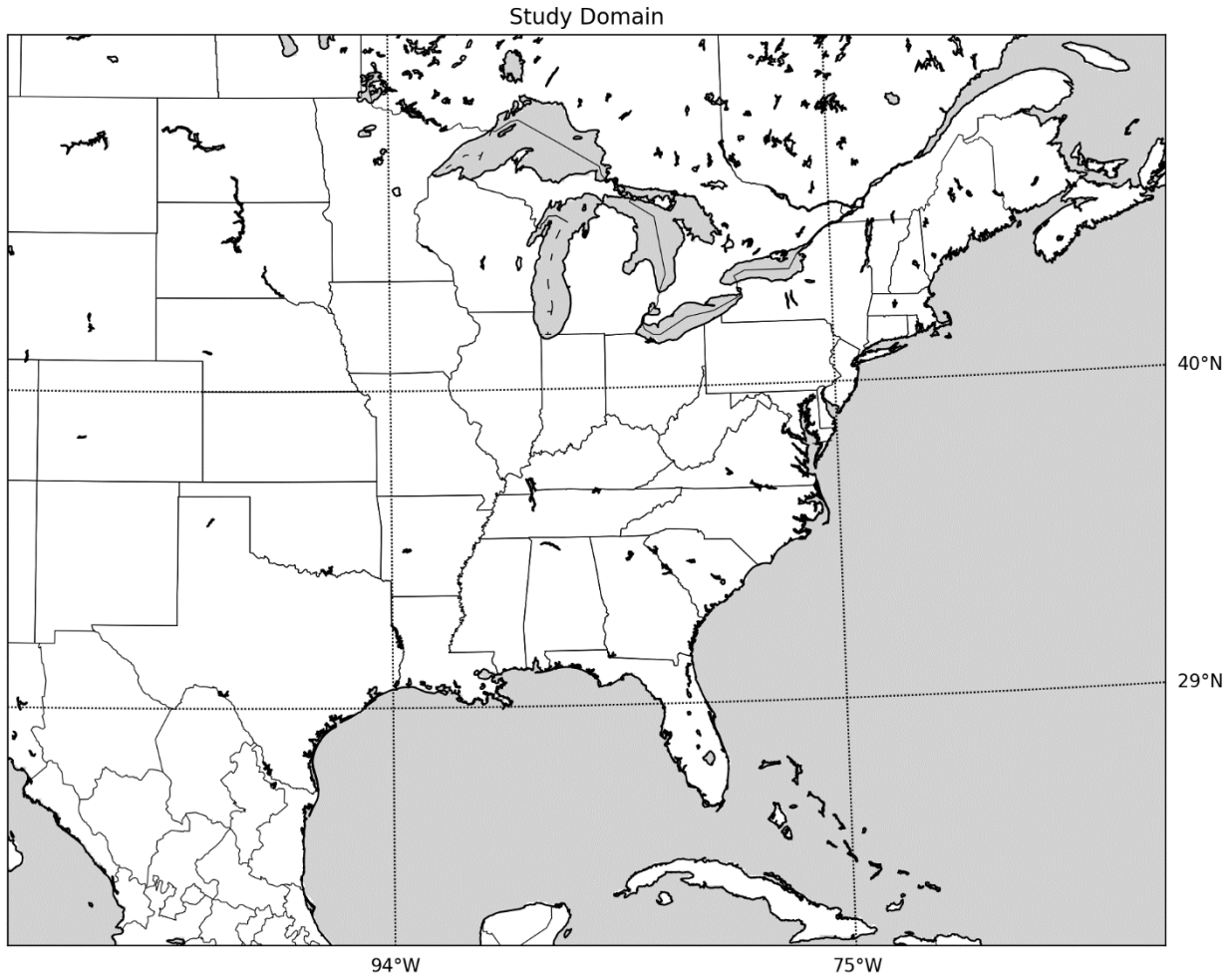

Figure 2.1. The study domain is contained within the box demarcated by the two parallels, 29°N and 40°N and the two meridians, 94°W and 75°W. Areas over the Gulf of Mexico and Atlantic Ocean were not considered to be parts of the study domain and were not included in skill score calculations.

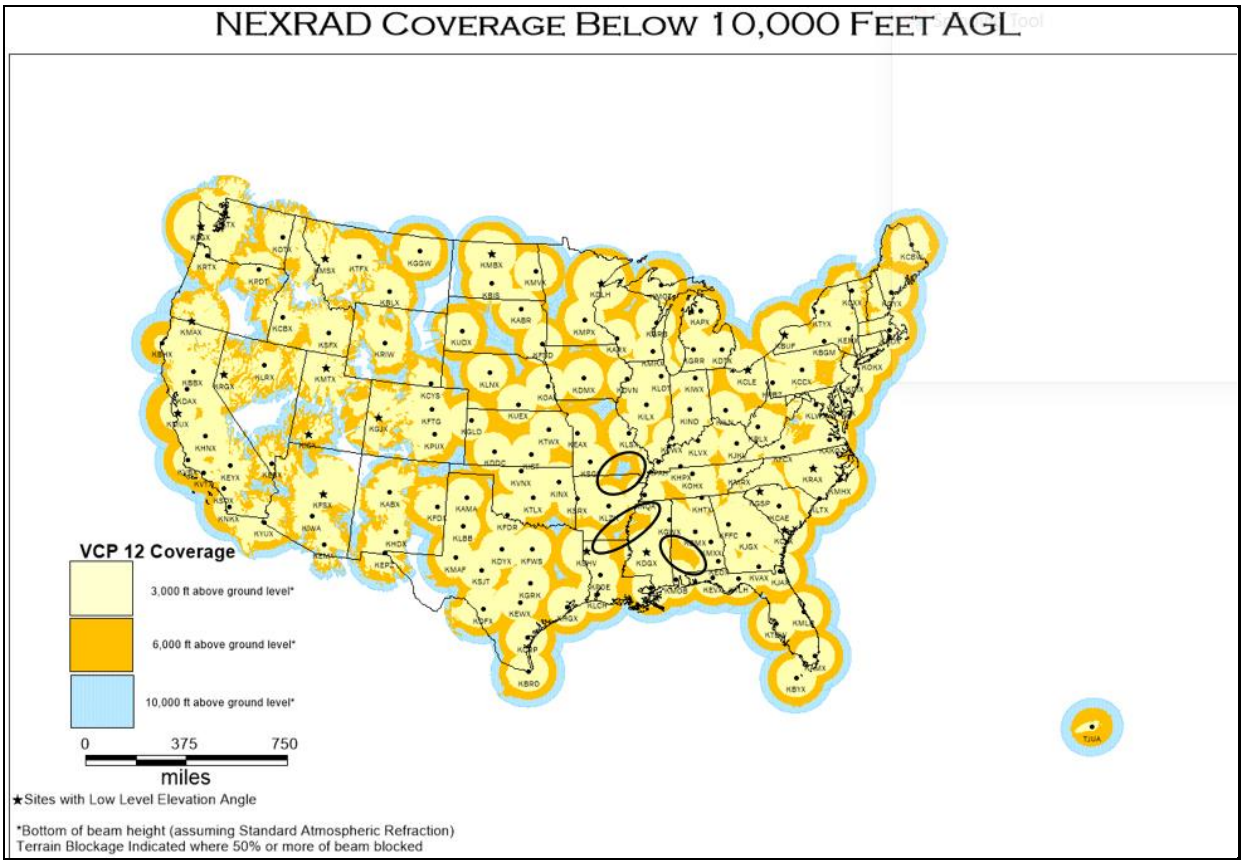

Figure 2.2. NEXRAD coverage by lowest beam height across the contiguous 48 states. Several regions within the study domain where the lowest beam height is above 6,000 feet are indicated by black ovals. Plot created by NOAA Radar Operations Center, https://www.roc.noaa.gov/WSR88D/.

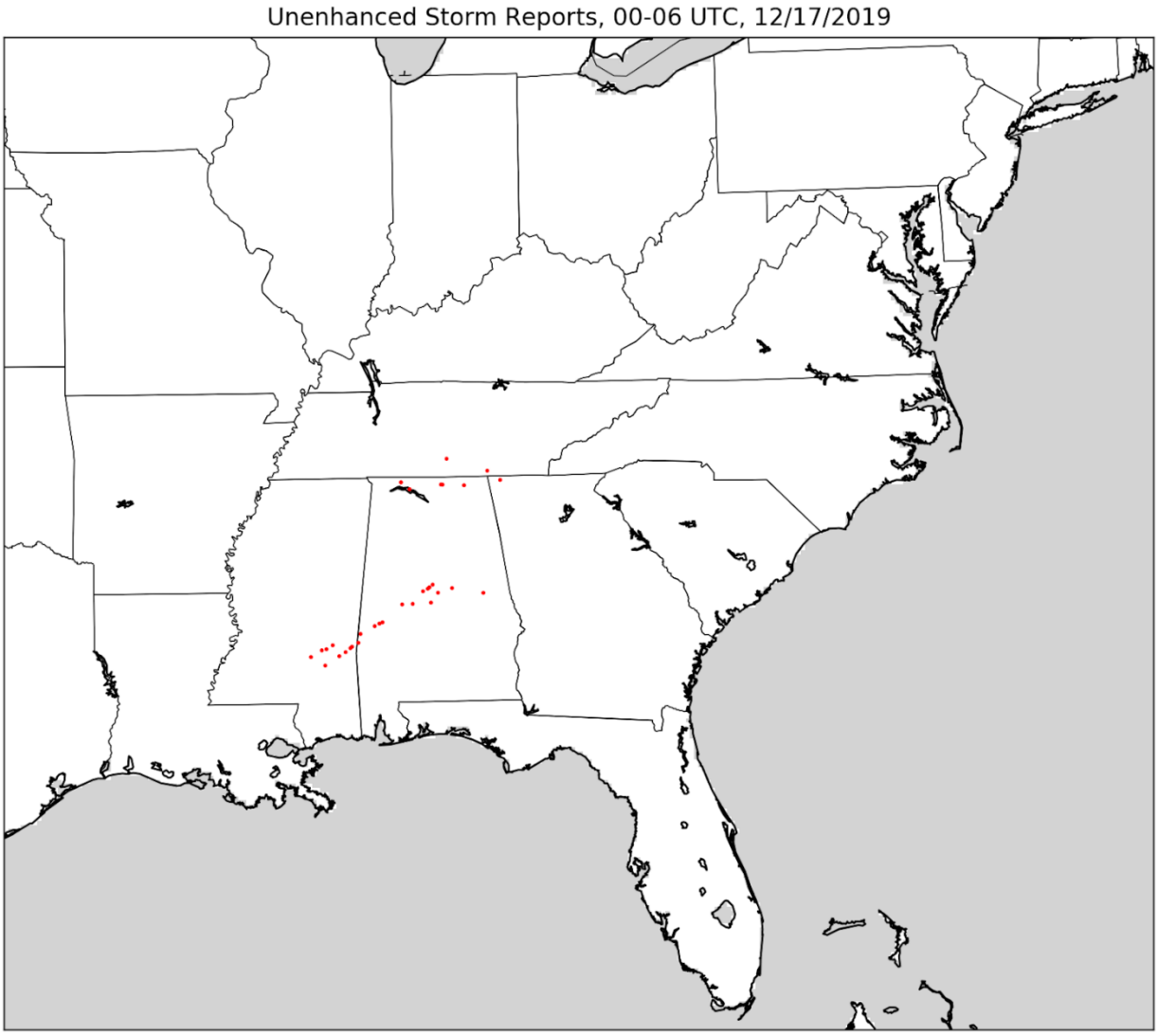

Figure 2.3a. Gridded LSR data (red dots) from 00-06 UTC 17 December 2019. This field was transferred onto the HREF grid using nearest-neighbor methods without any enhancement.

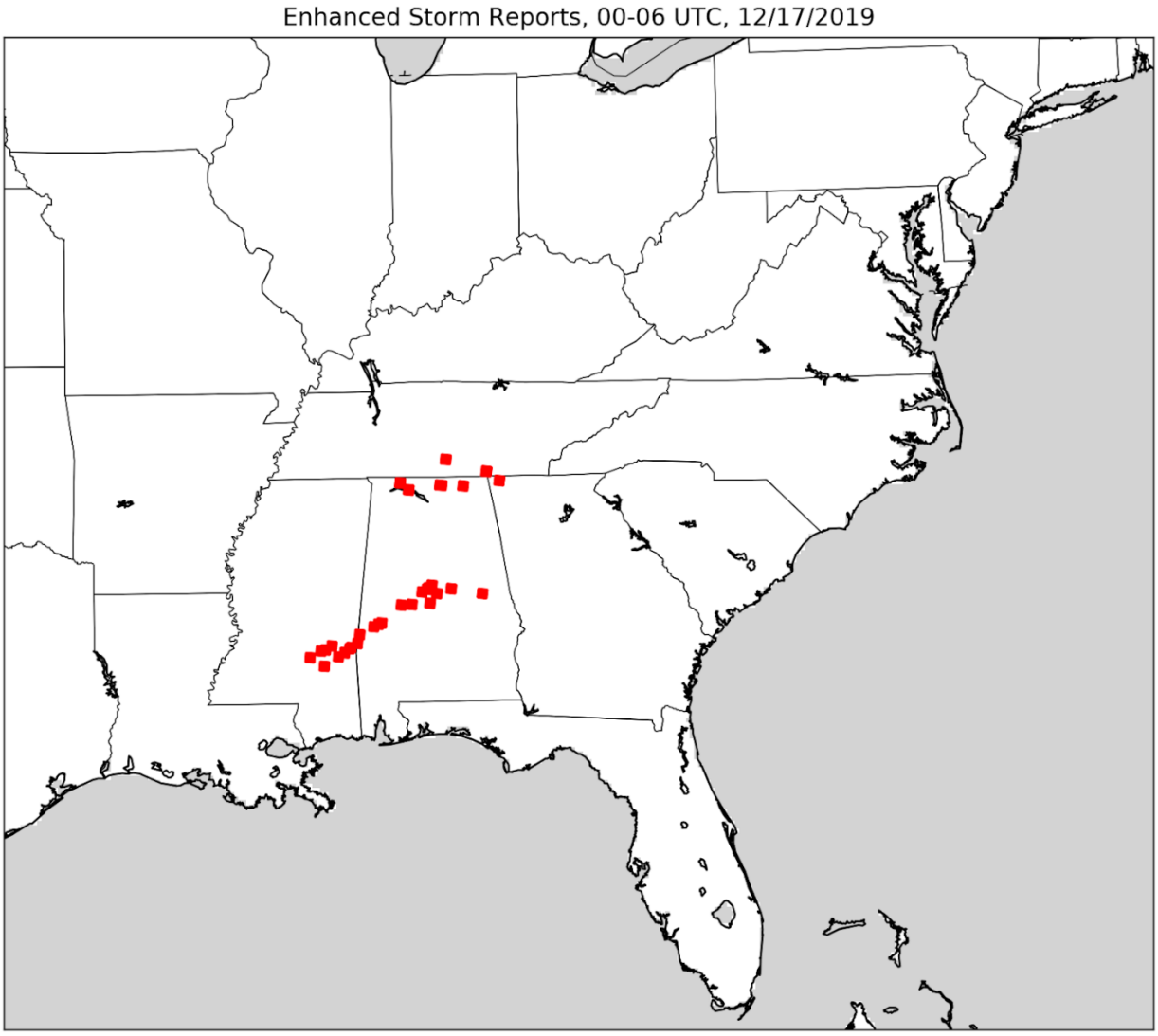

Figure 2.3b. As in Figure 2.3a, but the gridded LSR data has been enhanced such that adjacent grid cells in a 7x7 area around the grid cell which contained the LSR are treated as if they also contained an LSR.

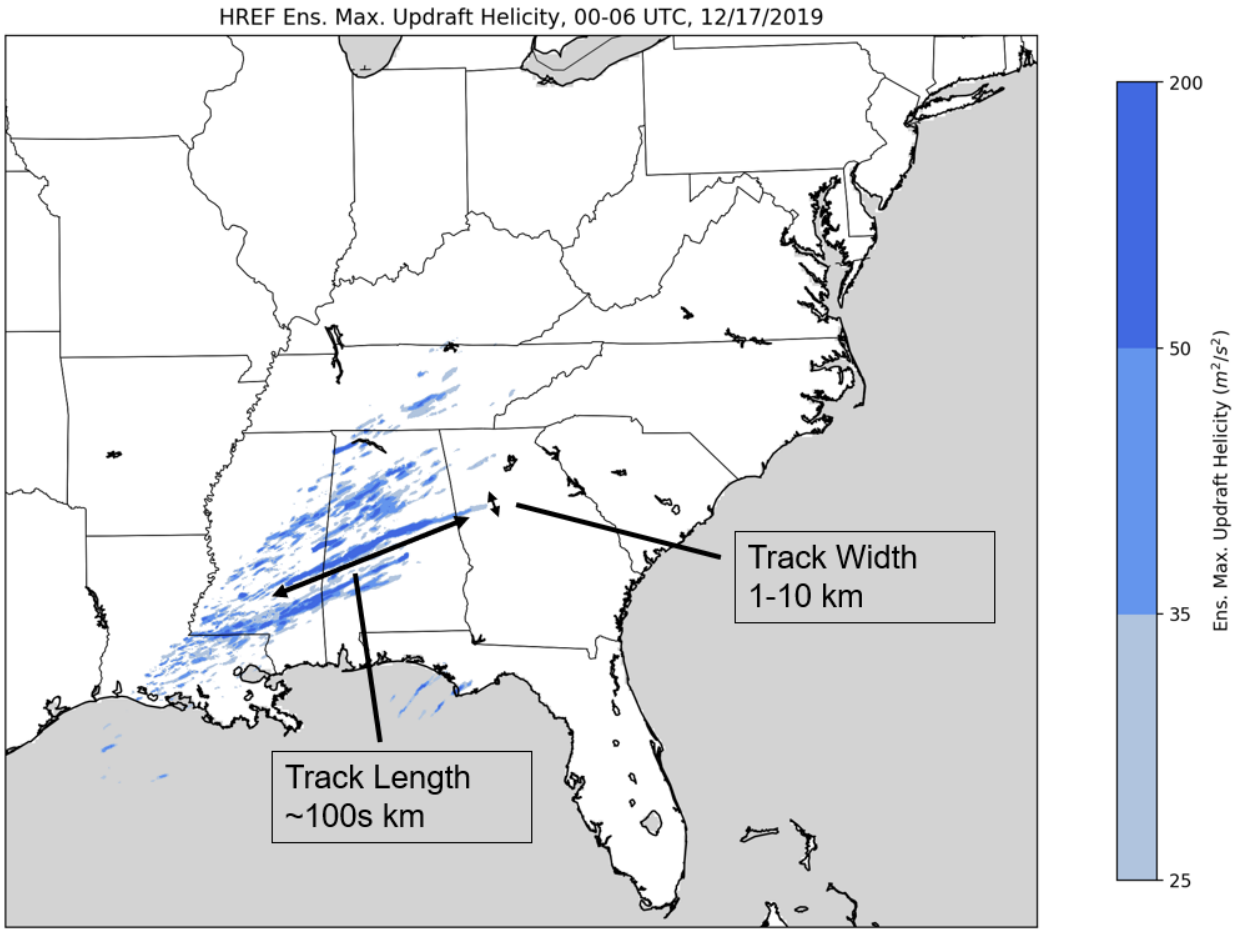

Figure 2.4. HREF Ensemble Maximum 0-3 km. Updraft Helicity (blue shading), 00-06 UTC, 17 December 2019. The disparity between the length and width of the UH swaths is denoted by the embedded text boxes.

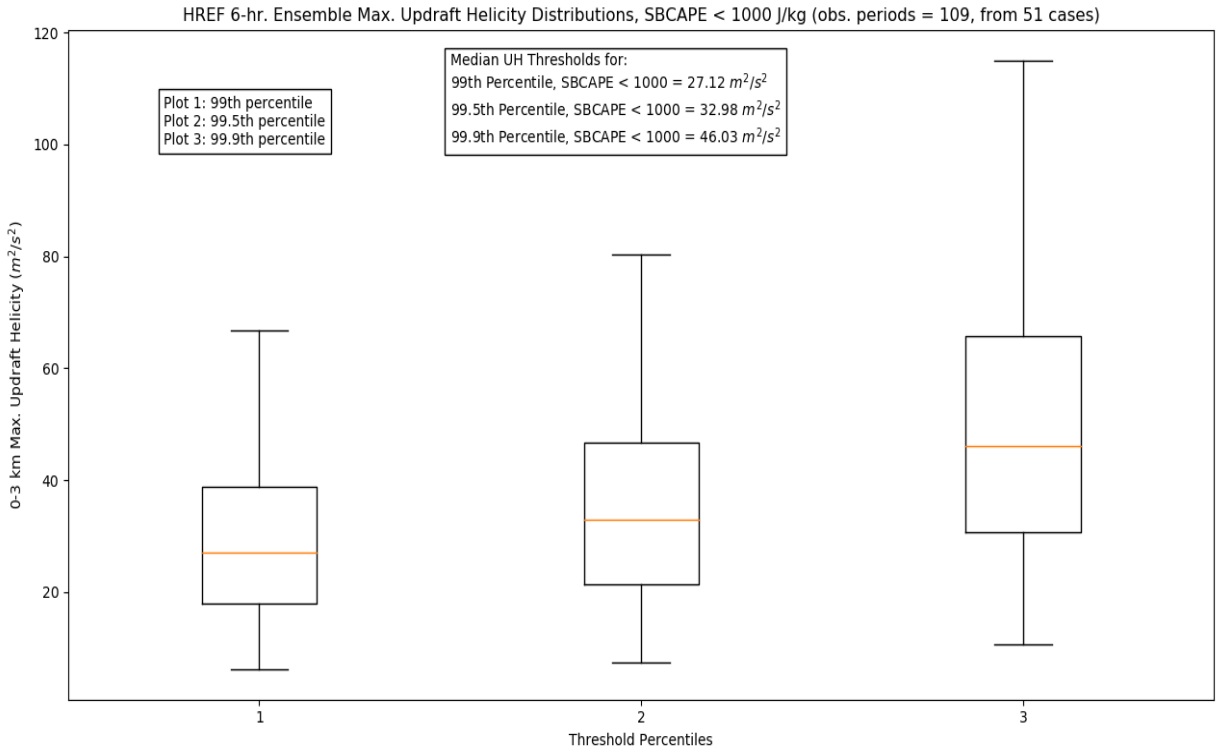

Figure 2.5. Distributions of the 99th (left), 99.5th (center), and 99.9th (right) percentile HREF six-hour Ensemble Maximum Updraft Helicity values in regions where SBCAPE < 1000 J/kg. The distributions are taken from 109 six-hour observation periods spanning 51 cases.

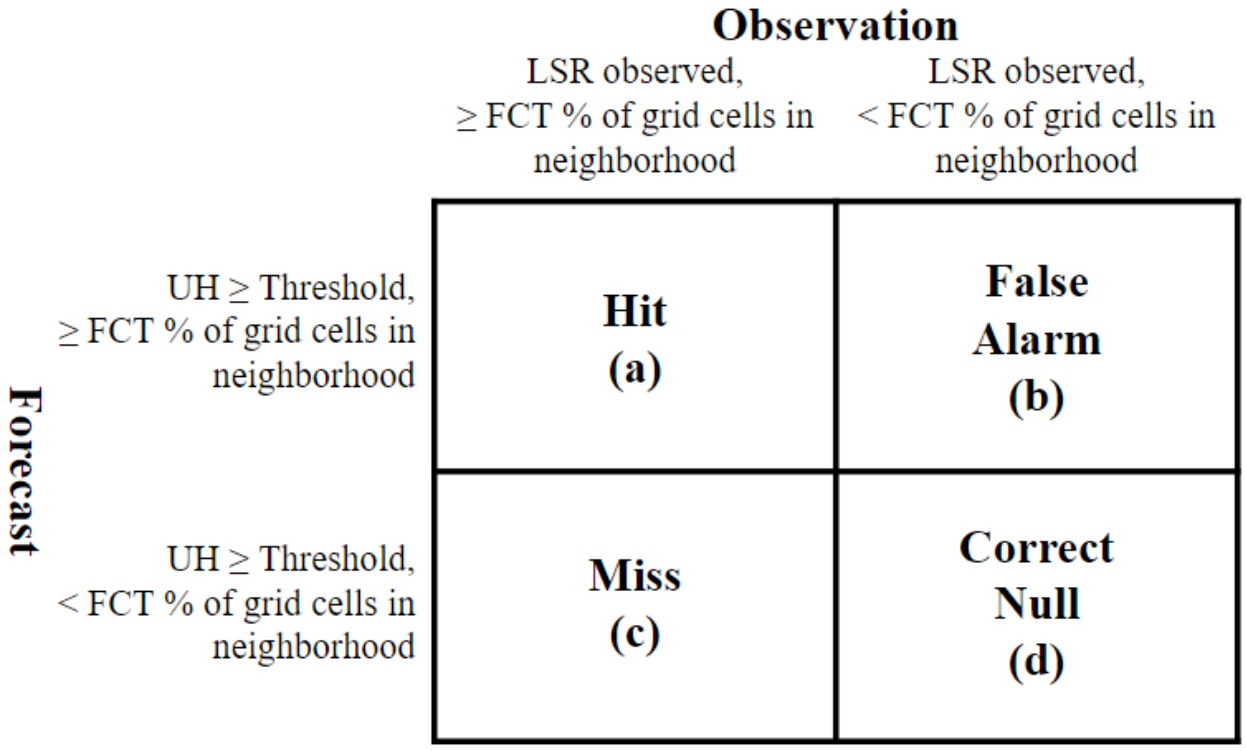

Figure 2.6. Contingency Table with definitions for how forecasts and observations were separated into "yes" and "no" bins. Both the Updraft Helicity Threshold ("Threshold") and Fractional Coverage Threshold (FCT) were manipulated for different tests.

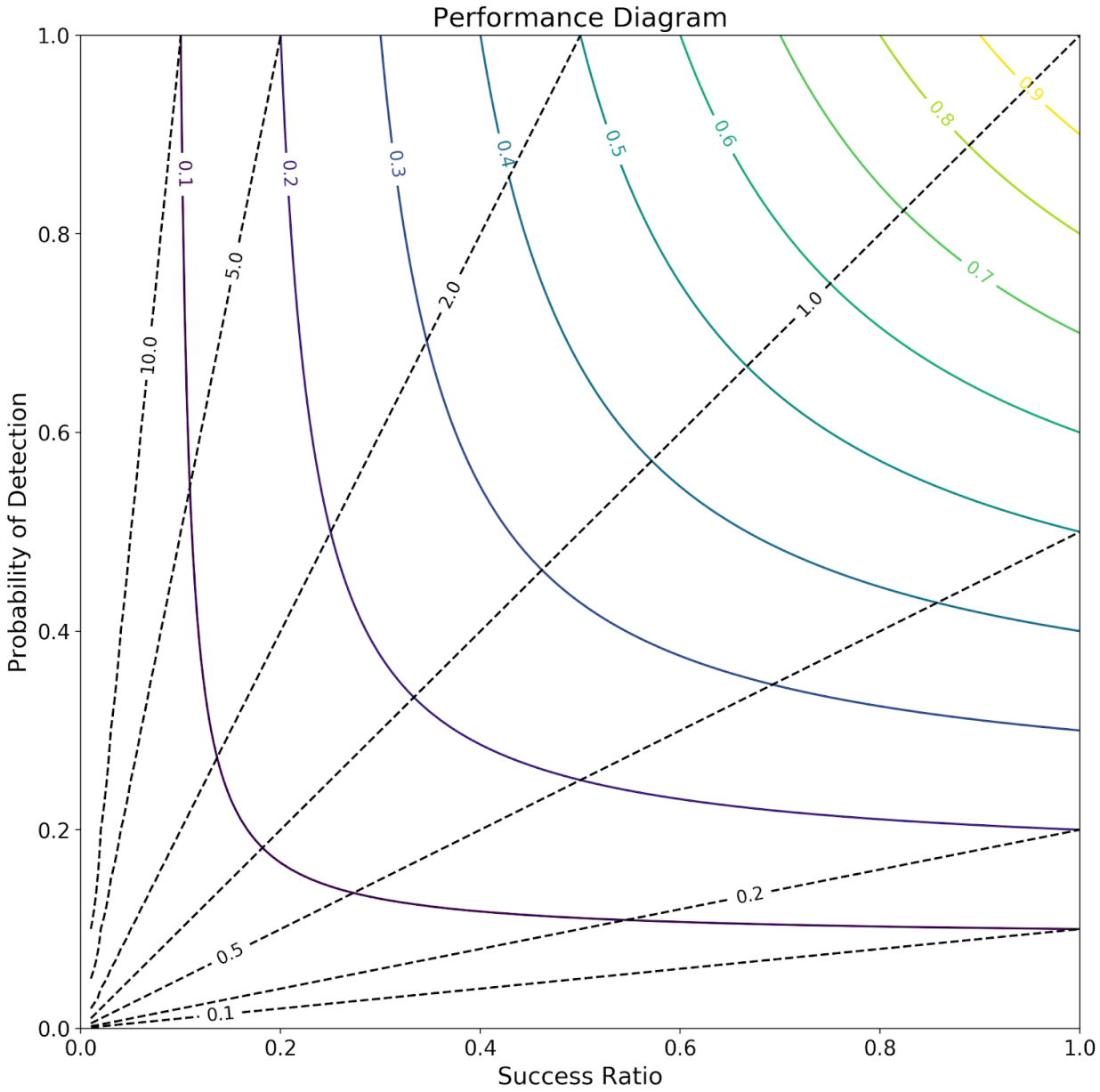

Figure 2.7. Performance Diagram as described in Roebber (2009). Success Ratio (SR) is plotted along the abscissa, while Probability of Detection (POD) is plotted along the ordinate. Critical Success Index (CSI) values are determined by a combination of SR and POD as described by (11), and CSI contours are shown by the multicolored solid lines. Forecast bias as described by (12) is contoured by the dashed black lines.

CHAPTER 3

Results

3.1 Fractions Skill Scores

3.1.1 Fractions Skill Scores for Selected Events

Before discussing the performance of HREF UH forecasts across the entire selection of events, which is done primarily through mean statistical values, it is beneficial to engage with a small selection of events to see what well, moderately well, and poorly forecasted individual events look like. In order to do this, plots with enhanced LSRs as described in Section 2 overlaid on HREF UH forecasts were created for three events: 00-06 UTC 17 December 2019, which represents a well forecasted event, 06-12 UTC 6 February 2020, which represents a moderately well forecasted event, and 00-06 UTC 23 January 2018, which represents a poorly forecasted event. The inclusion of the plots for these individual events will help to provide visual confirmation of the spatial distributions of forecasts and observations which are representative of varying degrees of forecast skill.

Figure 3.1 shows the spatial distribution of UH swaths and LSRs for the 00-06 UTC 17 December 2019 event. The plots were created such that values of UH < 25 m^2/s^2 are left unshaded, and that shading color changes to progressively darker shades of blue as UH values exceed progressively higher UH thresholds. For this event, nearly all LSRs are colocated with the center of the strongest UH swaths. Additionally, there is generally little UH activity outside of the regions where LSRs are observed with the exception of some locations in Louisiana and central Mississippi where swaths of lower UH values (but exceeding the lowest set UH thresholds) exist. The variation in Fractions Skill Scores (FSS) by UH threshold confirms expectations based on Figure 3.1: FSS for UH ≥ 25 m^2/s^2, which allows for the most spatially expansive swaths of UH and therefore the most overforecasting, is lowest at 0.23; FSS increases as UH threshold increases and UH swaths become more spatially limited (FSS for UH ≥ 35 m^2/s^2 is 0.4, FSS for UH ≥ 50 m^2/s^2 is 0.69) before ultimately decreasing for the highest UH threshold for which UH swaths exceeding the threshold are so spatially limited as to change the sign of the forecasting bias to an underforecast (FSS for UH ≥ 75 m^2/s^2 is 0.4).

In comparison to the first event, it is evident from Figure 3.2 that the UH forecasts for the 06-12 UTC 6 February 2020 event are somewhat less skillful. The main region of UH activity is shifted slightly to the south and east of the main region where LSRs are observed, although the neighborhood verification methods will forgive some of this spatial error. In addition, areas of UH $\geq$ 75 m^2/s^2 are almost non-existent for this case, which would suggest poorer forecast skill for this UH threshold given the total amount of LSRs. Once again, FSS values largely confirm visual adjudications of skill, as values are lower for all UH thresholds compared to the first event (FSS for UH $\geq$ 25 m^2/s^2, 35 m^2/s^2, and 50 m^2/s^2 is 0.09, 0.18, and 0.13, respectively), and the FSS value for UH $\geq$ 75 m^2/s^2 is the lowest of all thresholds, at 0.01.

For the last of the selected cases, the skill values for the 00-06 UTC 23 January 2018 event can be easily interpreted from Figure 3.3. This is one of a few examples of a complete miss by the HREF UH forecast, as no values of UH are forecasted above any of the UH thresholds in the area where LSRs are observed. As a result, FSS values are 0 for all UH thresholds, confirming the forecast miss and indicating that the HREF UH forecast has no skill for this event.

3.1.2 Aggregated Fractions Skill Scores

Considering that the verification process was conducted for 144 separate events, it would be impractical to present findings on an event-by-event basis; additionally, little information about the overall performance of the HREF can be gathered when looking at events individually. As a result, the findings from this study are presented in an aggregated fashion, averaging statistical performance metrics over the entire group of events. Analysis of overall HREF UH performance using the FSS is done using two different averaging methods. The first averaging method is conducted by considering each of the 144 events to be equal in importance and weight; in other words, an unweighted mean FSS value ($\overline{FSS}$) is calculated:

$$\overline{FSS} = \frac{1}{n}\sum_{i=1}^{n} FSS_i \qquad (3.1)$$

where n is the total number of events (n = 144) and FSS_i is the FSS value for the ith event. In addition to the unweighted mean FSS value, an additional averaging method is used that assigns weights to each event based on the amount of combined tornado and damaging wind LSRs that the event has. The weighted mean FSS value ($\overline{FSS_{weighted}}$) is calculated as follows:

$$\overline{FSS_{weighted}} = \sum_{i=1}^{n} \frac{LSR_i}{\text{Total LSRs}} \times FSS_i \qquad (3.2)$$

where n is again the total number of events, LSR_i is the number of combined tornado and damaging wind LSRs for the ith event, Total LSRs is a constant representing the total number of combined tornado and damaging wind reports across all events (*Total LSRs* = 3148), and FSS_i again represents the FSS value for the ith event. Since the number of combined LSRs varies widely from event to event, so also does the weighting per event vary widely. For example, the minimum number of combined LSRs for an event is 1, which was achieved by 14 events, while the maximum number of combined LSRs for an event is 247, which occurred in the 11 Jan 2020 18-00 UTC event. As a result, each of the events which had 1 LSR received a weighting of 1/3148 (~.03%), while the 11 Jan 2020 18-00 UTC event received a weighting of 247/3148 (~7.8%). Consequently, the single 11 Jan 2020 18-00 UTC event has roughly 18.5 times the weight of the 14 events with a single LSR combined. The purpose of using two different averaging methods is not to suggest that one averaging method is superior or even preferable to the other. In fact, the purpose of using both an unweighted and a weighted mean FSS value is to highlight differences in skill based on the amount of reported severe activity that a particular event has. Admittedly, forecasters would be highly unlikely to know the amount of severe reports that an event would have before its commencement. However, differences between unweighted and weighted mean performance metrics may reveal information about variations in skill of the HREF UH forecasts for events which occur on different spatial and temporal scales, since events on larger scales (e.g., long-lasting QLCS events) would more likely be associated with large amounts of LSRs than events which occur on smaller scales (e.g., isolated, short-lived convection). Ultimately, it is valuable to represent aggregated performance metrics in these two ways to show if HREF UH forecasts perform differently in widespread, outbreak-type events compared to marginal events which are generally more difficult to forecast.

3.1.3 Variations in Fractions Skill Scores by UH Threshold

While the unweighted and weighted average FSS values are one useful method for comparing and contrasting the overall performance of HREF UH forecasts by UH threshold, another method which can be employed is aggregating counts of which UH threshold had the best and worst performance (i.e., highest and lowest FSS value, respectively) over all 144 events.

Figure 3.4 shows that the lowest threshold used, UH ≥ 25 m^2/s^2, had the most skill for the greatest number of events, with the higher UH thresholds having the most skill for successively fewer events. 16 events (~11% of the total number of events) had no UH threshold which had the highest skill, which was exclusively due to the fact that the FSS for all UH thresholds in these events was 0. For events which did have non-zero FSS values, if the maximum FSS across all UH thresholds occurred for either UH ≥ 25 m^2/s^2 or 75 m^2/s^2, FSS values would either monotonically decrease (if UH ≥ 25 m^2/s^2 had the maximum FSS) or monotonically increase (if UH ≥ 75 m^2/s^2 had the maximum FSS) as UH thresholds increased. If an intermediate UH threshold (UH ≥ 35 m^2/s^2 or 50 m^2/s^2) had the maximum FSS, FSS values would increase as UH thresholds increased up until the UH threshold where the maximum FSS occurred, and then decrease subsequently as UH thresholds increased. In contrast to Figure 3.4, Figure 3.5 displays the aggregated counts of which UH thresholds recorded the lowest skill across the entire sample of events. For the vast majority of events (~85%), either UH ≥ 75 m^2/s^2 or multiple UH thresholds recorded the poorest FSS values, while UH ≥ 25 m^2/s^2 recorded the poorest FSS value much less often (~15%) and the intermediate UH thresholds never had the poorest FSS value. Almost invariably, if UH ≥ 75 m^2/s^2 had the lowest skill in an event, it would be because there would be little or no areas of UH exceeding the threshold anywhere in the domain, while there were extensive swaths of UH exceeding lower thresholds. If UH ≥ 25 m^2/s^2 was the poorest-performing threshold for a given event, it was typically because the swaths of UH exceeding the threshold were so expansive that there would be an "overforecast" of severe convection by the model, while the more stringent higher UH thresholds would more closely match the scale of observed severe convection, and thus exhibit higher levels of skill. Finally, for the events where multiple UH thresholds had the lowest attributed FSS, either a lower UH threshold exhibited maximum skill and UH swaths above the higher thresholds were non-existent (37/53 events) or all UH thresholds exhibited no skill (16/53 events).

Comparing the unweighted average FSS values for the various UH thresholds across all events, the lower three UH thresholds have fairly similar average skill values, ranging between $\overline{FSS}$ = 0.132 (UH ≥ 50 m^2/s^2) and $\overline{FSS}$ = 0.152 (UH ≥ 35 m^2/s^2) (Figure 3.6). UH ≥ 75 m^2/s^2 is the least skillful threshold by a considerable margin ($\overline{FSS}$ = 0.059), exhibiting less than half the amount of average skill as the other three UH thresholds. It is notable that despite having the highest FSS value in a plurality of events, UH ≥ 25 m^2/s^2 is less skillful on average ($\overline{FSS}$ =

0.138) than UH ≥ 35 m^2/s^2 using the unweighted mean. This would suggest that while UH ≥ 25 m^2/s^2 more frequently possesses the highest skill, UH ≥ 35 m^2/s^2 may be a more consistently skillful UH threshold. Nevertheless, considering the range of FSS values between a forecast with no skill (FSS = 0) and a perfect forecast (FSS = 1), when using the verification methodology of this study, all UH thresholds perform generally poorly on average.

When average FSS values are calculated by weighting events based on number of combined tornado and damaging wind LSRs, improvements in average skill are experienced for all UH thresholds, with mean FSS values roughly increasing by a factor of 2 (Figure 3.7). Relatively small differences in the amount of improvement between UH thresholds allows for UH ≥ 25 m^2/s^2 ($\overline{FSS_{weighted}}$ = 0.276) to draw level with UH ≥ 35 m^2/s^2 ($\overline{FSS_{weighted}}$ = 0.275) in terms of being the most skillful UH threshold on average. Additionally, using the weighted average, the gap between the average skill of UH ≥ 50 m^2/s^2 ($\overline{FSS_{weighted}}$ = 0.219) and the lower two UH thresholds increases, although this threshold is still clearly more skillful than the highest UH threshold, UH ≥ 75 m^2/s^2 ($\overline{FSS_{weighted}}$ = 0.111). Given that average skill improved for all UH thresholds using a weighted mean based on amount of LSRs, it follows that HREF UH forecasts are more skillful for events which contain greater amounts of LSRs across all thresholds, with slight variations in the amount of improvement in performance from threshold to threshold.

3.1.4 Variation in Fractions Skill Scores by LSR Type

The results presented thus far explored differences in average skill using different forecast quantities while keeping the observation dataset constant. Weighted and unweighted mean FSS values in Figures 3.6 and 3.7 were calculated by verifying HREF UH forecasts against a combined dataset of tornado and damaging wind LSRs. However, it is also important to observe how well HREF UH forecasts perform against tornado LSRs and damaging wind LSRs individually to see if UH forecasts are more skillful with one LSR type or another. Using an unweighted average, there is a consistent pattern across all UH thresholds where mean FSS values when verifying against only damaging wind LSRs are similar to mean FSS values when verifying against all LSRs (Figure 3.8). In contrast, mean FSS values when verifying against only tornado LSRs are substantially lower, although the gap between mean FSS values for tornadoes and damaging winds appears to shrink as UH threshold increases. It appears that the

shrinking of the gap in performance between damaging wind LSRs and tornado LSRs at higher thresholds is a result of the difference between the UH threshold at which different LSR types show greatest skill, with peak mean FSS values occurring at UH ≥ 35 m^2/s^2 for damaging wind LSRs and UH ≥ 50 m^2/s^2 for tornado LSRs. Figure 3.9 shows how increases in mean FSS using an LSR-weighting method are consistent across all observation types at all thresholds. In addition, the pattern where the gap between performance against only damaging wind LSRs and only tornado LSRs decreases with increasing UH threshold is more pronounced using LSR-weighted mean methods. Using this method, average FSS values for damaging wind LSRs are maximized both for UH ≥ 25 m^2/s^2 and 35 m^2/s^2, while average FSS values for tornado LSRs are maximized at UH ≥ 50 m^2/s^2.

3.1.5 Variations in Fractions Skill Scores by Time of Day

Because the six hour observation windows are standardized for all 144 events, it is possible to group together events which occur during the same window of time and compare and contrast average skill values for events which occur at different times of the day. Fortunately, the 144 events were spread relatively evenly across each of the four six hour time frames, with 30 events occurring in the 12-18 UTC time frame, 47 events occurring in the 18-00 UTC time frame, 39 events occurring in the 00-06 UTC time frame, and 28 events occurring in the 06-12 UTC time frame. Figure 3.10 reveals that the relative performance of the different UH thresholds varies across the four time windows, with UH ≥ 25 m^2/s^2 recording the highest mean FSS value for the 18-00 UTC time period, UH ≥ 35 m^2/s^2 recording the highest mean FSS value for the 12-18 UTC and 06-12 UTC time periods, and UH ≥ 50 m^2/s^2 recording the highest mean FSS value for the 00-06 UTC time period. Nevertheless, all three of the lower UH thresholds are relatively close in mean FSS values for all time periods, with UH ≥ 75 m^2/s^2 consistently lagging behind. When calculating the mean FSS values using the LSR-based weighting method, all thresholds improve in a consistent manner with overall improvements (Figure 3.6) as well as improvements when isolating by LSR type (Figure 3.8). As was the case with the unweighted mean FSS values, the highest performing UH threshold rotates amongst the lowest three thresholds during different portions of the diurnal cycle, with UH ≥ 25 m^2/s^2 having the highest mean FSS value during the 18-00 UTC time period, UH ≥ 35 m^2/s^2 having the highest mean FSS value during the 12-18 UTC time period, UH ≥ 50 m^2/s^2 having the highest mean FSS value during the 06-12 UTC time

period, and both UH ≥ 25 m^2/s^2 and 35 m^2/s^2 sharing the highest mean FSS value for the 00-06 UTC time period. Likewise, UH ≥ 75 m^2/s^2 also recorded the lowest mean FSS values for all four six-hour time windows, although the gap in skill between the lower three UH thresholds and UH ≥ 75 m^2/s^2 was smaller using the LSR-weighted mean, especially for the 06-12 UTC time frame. In general, the relationship between mean FSS values for the various UH thresholds across different portions of the day follows the overall trend, with UH ≥ 25 m^2/s^2, 35 m^2/s^2, and 50 m^2/s^2 performing roughly equally and UH ≥ 75 m^2/s^2 performing less well.

3.2 Contingency Table Statistics

As was the case for the FSS results, the main focus of the contingency table statistics and performance diagrams will be on the average level of skill across all 144 events. In contrast with the FSS results, individual events can be plotted onto a performance diagram to give a greater sense of the distribution of skill across all events. Therefore, for the performance diagram plots (Figures 3.12-3.20), each of the 144 events is represented by a small dot, colored either based on UH threshold or time of day, while the mean values for each UH threshold or time of day are represented by larger dots. It is important to note that while the parameter space of POD and FAR values (and by extension, SR values) typically lies between 0 and 1, there are occasions when the extent of observations or forecasts above a specified threshold never exceeds the fractional coverage threshold set for an event. In this case, there would be a 0 in the denominator of POD and SR calculations, which causes the calculations to break down and no longer produce useful solutions. This occurs more frequently for higher UH thresholds, which are usually more spatially confined, and for higher fractional coverage thresholds. In order to avoid having to exclude a substantial number of events, whenever an event has no hits or misses (leading to a 0 in the denominator of the POD calculation), the POD for that event is considered to be 0, as is suggested by Jolliffe and Stephenson (2003). Likewise, if an event has neither hits nor false alarms (leading to a 0 in the denominator of the FAR calculation), the FAR for that event is considered to be 0, and the SR is considered to be 1.

There are two different methods which were used to calculate the mean skill values for the contingency table statistics. As in the FSS mean calculations, the first method for calculating the mean contingency table statistics values involves valuing all events equally, irrespective of

the amount of LSRs which occurred in the events. Mean POD ($\overline{POD}$) and SR ($\overline{SR}$) values using this "unweighted" method are calculated as follows:

$$\overline{POD} = \frac{1}{n}\sum_{i=1}^{n} POD_i \qquad (3.3)$$

$$\overline{SR} = \frac{1}{n}\sum_{i=1}^{n} (1-FAR)_i \qquad (3.4)$$

where n is the total number of events (n = 144), POD_i is the probability of detection for the ith event and $(1\text{-}FAR)_i$ is the success rate for the ith event (SR = 1-FAR). Ultimately, the unweighted mean POD and SR values are combined in order to calculate the unweighted mean CSI value ($\overline{CSI}$) in the following manner:

$$\overline{CSI} = \frac{1}{\frac{1}{\overline{SR}} + \frac{1}{\overline{POD}} - 1} \qquad (3.5)$$

In order to assess the differences in performance of HREF UH forecasts in events with high amounts of combined tornado and damaging wind LSRs and events with few LSRs, an LSR-based weighted mean is also calculated for the various contingency table statistics. LSR-weighted mean POD ($\overline{POD_{weighted}}$) and SR ($\overline{SR_{weighted}}$) values are calculated in a similar fashion to the LSR-weighted mean FSS values:

$$\overline{POD_{weighted}} = \sum_{i=1}^{n}\left(\frac{LSR_i}{\text{Total LSRs}} \times POD_i\right) \qquad (3.6)$$

$$\overline{SR_{weighted}} = \sum_{i=1}^{n}\left(\frac{LSR_i}{\text{Total LSRs}} \times (1-FAR)_i\right) \qquad (3.7)$$

where n, LSR_i, and Total LSRs have the same meaning and associated values as in (3.2) and POD_i and $(1\text{-}FAR)_i$ have the same meaning as in (3.3) and (3.4), respectively. Combining the weighted mean POD and SR values together in the same manner as in (3.5), the weighted mean CSI is calculated as follows:

$$\overline{CSI_{weighted}} = \frac{1}{\frac{1}{\overline{SR_{weighted}}} + \frac{1}{\overline{POD_{weighted}}} - 1} \qquad (3.8)$$

In the performance diagrams, weighted and unweighted mean values are shown together with differing representations (black-outlined dots vs. grey-outlined dots) in order to make a direct comparison between the two mean values.

3.2.1 Variations in Contingency Table Statistics by UH Threshold

As was discussed in Section 2, in order to use neighborhood verification methods with contingency table statistics, a fractional coverage threshold must be set and exceeded by both the forecast and observation dataset in order to register as a "yes" forecast or observation. In this analysis, the fractional coverage threshold is varied, starting with the minimum possible threshold (1/625 grid cells in a neighborhood, ~0.1%) and increasing up to 10%, with intermediate fractional coverage thresholds also examined (1%, 2%, 5%). Fractional coverage thresholds are set at low percentages because both forecast UH swaths and observations occupy only a spatially-limited portion of the domain. Despite the fractional coverage threshold being one of the manipulated variables in this analysis, the main points to emphasize in this section are the differences in average skill between the various UH thresholds and the differences between the LSR-weighted and unweighted averages at the same UH threshold.

For the minimum fractional coverage threshold, lower UH thresholds exhibit low SR values (i.e., a high false alarm rate), while higher UH thresholds have relatively higher SR values. This is exemplified by the mean (SR, POD) ordered pairs showing up on the left hand side of the performance diagram for UH $\geq$ 25 m^2/s^2 and 35 m^2/s^2 before shifting over to the middle and eventually the right side of the diagram for UH $\geq$ 75 m^2/s^2 (Figure 3.12). As is inferred from the visual interpretation, unweighted mean SR values improve as UH threshold increases, from $\overline{SR} = 0.21$ for UH $\geq$ 25 m^2/s^2 up to $\overline{SR} = 0.65$ for UH $\geq$ 75 m^2/s^2. This is not an unexpected pattern, as fewer false alarms (a higher success rate) would be expected on average for the more rigorous and spatially-limited higher UH thresholds. Another trend which is stronger in magnitude is that unweighted mean POD values are maximized at UH $\geq$ 25 m^2/s^2 ($\overline{POD} = 0.67$) and decrease consistently as UH threshold increases up to UH $\geq$ 75 m^2/s^2 ($\overline{POD} = 0.13$). Once again, this is not particularly surprising, as more forecast misses would be expected at higher UH thresholds. Mainly as a result of poor mean SR values (although low mean POD values also adversely contribute for the higher UH thresholds), unweighted mean CSI values are fairly poor for all UH thresholds. While UH $\geq$ 75 m^2/s^2 clearly produces the lowest mean skill

value ($\overline{CSI}$ = 0.12), it is somewhat competitive with the average skill values for the lower UH thresholds, with $\overline{CSI}$ = 0.19 for UH ≥ 25 m^2/s^2, $\overline{CSI}$ = 0.22 for UH ≥ 35 m^2/s^2, and $\overline{CSI}$ = 0.21 for UH ≥ 50 m^2/s^2.

When LSR-weighted mean statistics are calculated, skill metrics generally improve, with across the board improvements in POD and CSI and improvements in SR values for most UH thresholds. This suggests that there were fewer false alarms on average for the lower UH thresholds in events with high amounts of LSRs, a result which would not be unexpected. However, this result does not translate over for the highest UH threshold, where $\overline{SR_{weighted}}$ and $\overline{SR}$ values are roughly equal, suggesting that high activity events produce more or less the same amount of false alarms as low activity events. While the amount of improvement in SR values when using the LSR-weighted mean is variable across UH thresholds, POD values increase for all thresholds when using LSR-weighted mean methods.

As a result of improvements in $\overline{POD_{weighted}}$ values over $\overline{POD}$ values for all UH thresholds and $\overline{SR_{weighted}}$ values over $\overline{SR}$ values for lower UH thresholds, LSR-weighted mean CSI values are higher than unweighted mean CSI values across all UH thresholds. The separation between the highest and lowest mean CSI values increases slightly when the LSR-weighted mean methodology is employed. UH ≥ 25 m^2/s^2 and 35 m^2/s^2 are the most skillful UH thresholds according to the LSR-weighted mean CSI value ($\overline{CSI_{weighted}}$ = 0.37 for both thresholds), while $\overline{CSI_{weighted}}$ falls off slightly to 0.32 for UH ≥ 50 m^2/s^2, and quite substantially to 0.21 for UH ≥ 75 m^2/s^2. While mean POD values improved fairly consistently across all UH thresholds when using LSR weighting, it appears that the lack of improvement in mean SR values for the higher UH thresholds was the primary contributor to the increased separation between mean skill values when using the LSR-weighted mean methods. Additionally, this variation in mean SR improvement can also explain why the order of UH threshold by skill changes, as UH ≥ 25 m^2/s^2 goes from third-most skillful UH threshold to (tied) most skillful threshold and UH ≥ 50 m^2/s^2 goes from second-most skillful threshold to third-most skillful threshold.

Overall, as was the case for the FSS results, while there were substantial variations between mean skill values for the different UH thresholds, mean contingency table-based skill

values were, by and large, on the lower end of the scale. The benefit of the contingency table analysis comes from the revelation that the poor average skill displayed by HREF UH forecasts is due to high amounts of false alarms, especially for the lower UH thresholds, as well as high amounts of misses for the higher UH thresholds.

As fractional coverage threshold increases (Figures 3.13-3.16), there are a few notable trends in both the unweighted and weighted mean skill values. For the unweighted mean values, the most consistent pattern is a decrease in $\overline{POD}$ values as fractional coverage threshold increases among all UH thresholds. The trend in $\overline{SR}$ with increasing fractional coverage threshold is a bit more nuanced; $\overline{SR}$ values are fairly constant for the lower UH thresholds as fractional coverage threshold increases from the minimum to 5%, then increase rapidly as fractional coverage threshold increases from 2% to 10%. For the higher UH thresholds, $\overline{SR}$ values increase fairly consistently as fractional coverage threshold increases.

Since $\overline{POD}$ values generally decrease and $\overline{SR}$ values generally increase as fractional coverage threshold increases, the change in $\overline{CSI}$ values as fractional coverage threshold increases depends on the starting position of the mean (SR, POD) ordered pair within the parameter space of the performance diagram. For the minimum fractional coverage threshold, $\overline{POD}$ values are higher than $\overline{SR}$ values for the lower UH thresholds, so the decreases in $\overline{POD}$ values that occur as fractional coverage threshold increases are offset by improved $\overline{SR}$ values, and $\overline{CSI}$ values remain fairly constant. In contrast, because $\overline{POD}$ values are initially lower than $\overline{SR}$ values when using the minimum fractional coverage threshold for the higher UH thresholds, decreases in $\overline{POD}$ values as fractional coverage threshold increases force $\overline{CSI}$ values to decrease and are not offset by increases in $\overline{SR}$ values.

Some patterns which emerge as fractional coverage threshold increases in unweighted mean skill values are maintained when using the LSR-weighting method of calculating mean skill values. As was the case for the unweighted mean, weighted mean POD values decrease for all UH thresholds as fractional coverage threshold increases. In contrast to the trend in $\overline{POD_{weighted}}$ values, the trend in weighted mean SR values varies by UH threshold. The trend in $\overline{SR_{weighted}}$ values for the upper UH thresholds follows that which is observed in unweighted mean

SR values, although the magnitude of the increase in $\overline{SR_{weighted}}$ values as fractional coverage threshold increases is less pronounced, especially for UH ≥ 50 m^2/s^2. On the contrary, weighted mean SR values for the lower UH thresholds decrease as fractional coverage threshold increases. Finally, LSR-weighted mean CSI values decrease across all UH thresholds as fractional coverage threshold increases. UH ≥ 25 m^2/s^2 appears to have a slightly more resilient $\overline{CSI_{weighted}}$ value compared to the higher UH thresholds. The main contributor to the higher $\overline{CSI_{weighted}}$ values for UH ≥ 25 m^2/s^2 is its relatively high $\overline{POD_{weighted}}$ values, which would be expected given that UH ≥ 25 m^2/s^2 is the most spatially expansive forecast field and would be least affected by increasing the fractional coverage threshold.

3.2.2 Variations in Contingency Table Statistics by Time of Day

Variations in average HREF UH forecast skill can also be examined for different portions of the day, although sample sizes are admittedly much smaller. Equations (3) through (5) can be adapted to calculate unweighted mean POD, SR, and CSI values, while equations (6) through (8) can be adapted to calculate LSR-weighted mean POD, SR, and CSI values. The total number of events (*n*) will vary based on six-hour observation window: there are 30 events which fall into the 12-18 UTC time frame, 47 events which fall into the 18-00 UTC time frame, 39 events which fall into the 00-06 UTC time frame, and 28 events which fall into the 06-12 UTC time frame. For the LSR-weighted mean skill calculations, the *Total LSRs* quantity will also vary based on six-hour observation window. Out of a total of 3148 LSRs across all events, 586 occur in events between 12-18 UTC, 1067 occur in events between 18-00 UTC, 936 occur in events between 00-06 UTC, and 559 occur in events between 06-12 UTC. Since sample sizes for each six hour time frame are already small and increasing fractional coverage threshold decreases the number of usable events, only the minimum fractional coverage threshold is used for the contingency table statistics segregated by time of day.

Beginning with UH ≥ 25 m^2/s^2, all four six-hour windows have similar unweighted mean POD and SR values, and are located very close to each other on the performance diagram (Figure 17). The range of $\overline{POD}$ and $\overline{SR}$ values across the different six-hour windows is only .05. Since unweighted mean POD and SR values are so similar for all times of day, unweighted

mean CSI values are also very similar for each six-hour time frame. Using an LSR-weighted mean increases POD and SR values for all six-hour time frames, although the increase from $\overline{POD}$ to $\overline{POD_{weighted}}$ is considerably less for the 12-18 UTC time frame. Nevertheless, $\overline{CSI_{weighted}}$ values are considerably higher than $\overline{CSI}$ values for all time frames, and the range between highest and lowest CSI values remains small.

As UH threshold increases from 25 m^2/s^2 to 35 m^2/s^2, the spread between $\overline{POD}$ values grows substantially, while the spread between $\overline{SR}$ values increases only slightly (Figure 18). While unweighted mean POD values decrease for all six hour windows (as is consistent with overall mean POD value decreases from UH $\geq$ 25 m^2/s^2 to 35 m^2/s^2), the extent of the decrease varies substantially, resulting in the increased range of POD values. Ultimately, it appears that this discrepancy in the change in $\overline{POD}$ values is what primarily contributes to a slight increase in the range in $\overline{CSI}$ values for UH $\geq$ 35 m^2/s^2. Discrepancies in the amount of increase between unweighted and weighted mean POD, SR, and CSI values largely erase any increased variance in unweighted mean skill values, and $\overline{CSI_{weighted}}$ values continue to be roughly equal across six-hour time windows for UH $\geq$ 35 m^2/s^2.

For the highest two UH thresholds (Figures 19-20), it is important to note that even when using the minimum fractional coverage threshold, there are many events which do not contain UH values above the upper UH thresholds, which causes errors in calculating FAR values (and by extension, SR values). The impact of automatically assigning an SR value of 1 for these events will be progressively more impactful on mean skill values, especially considering that the subset of events in each six hour time frame was already fairly small. With that being said, trends which were observed as UH threshold increases across the entire set of events are also observed in each of the four six hour time windows. Improvements in skill between $\overline{CSI}$ and $\overline{CSI_{weighted}}$ values are caused more and more by improvements between $\overline{POD}$ and $\overline{POD_{weighted}}$ values, as $\overline{SR_{weighted}}$ values are roughly equal to or lower than $\overline{SR}$ values for some time windows, particularly for UH $\geq$ 75 m^2/s^2. For both UH $\geq$ 50 m^2/s^2 and 75 m^2/s^2, the greatest increase in mean skill between $\overline{CSI}$ and $\overline{CSI_{weighted}}$ values occurs for the 06-12 UTC time period, and the improvement between unweighted and weighted mean values is almost exclusively due to

$\overline{POD_{weighted}}$ being considerably greater than $\overline{POD}$. Considering that the 06-12 UTC time period also had the greatest or second-greatest improvement between $\overline{CSI}$ and $\overline{CSI_{weighted}}$ values for the lower UH thresholds, this could suggest that HREF UH forecasts perform better for high activity events during the overnight period, although diurnal variations in LSR reporting biases and small sample sizes could also explain the improvements in mean skill.

Overall, these results suggest that differences in HREF UH forecast skill are quite insignificant across the four different six hour time periods, albeit with a fairly limited sample of events. Mean skill values for the different time windows roughly match each other, and the range in $\overline{CSI}$ and $\overline{CSI_{weighted}}$ values across time periods remains fairly small for the lower UH thresholds where there is a larger sample of events that do not encounter issues with non-real FAR values. A particularly surprising finding which spans all UH thresholds is that mean unweighted SR for 18-00 UTC is the worst of all the time periods for three of the four UH thresholds. This is contrary to what would be expected, considering that 18-00 UTC within the study domain represents the afternoon and evening hours, when false alarms due to reporting biases would likely be at their lowest. Irrespective of variations in skill between different times of day and mean weighting methods, it is important to reiterate that all mean skill values presented in this section are generally poor. Notwithstanding issues with sample size and general applicability, the skill score results by time of day (with some variation) broadly align with the overall skill score results presented in Section 3.2.1.

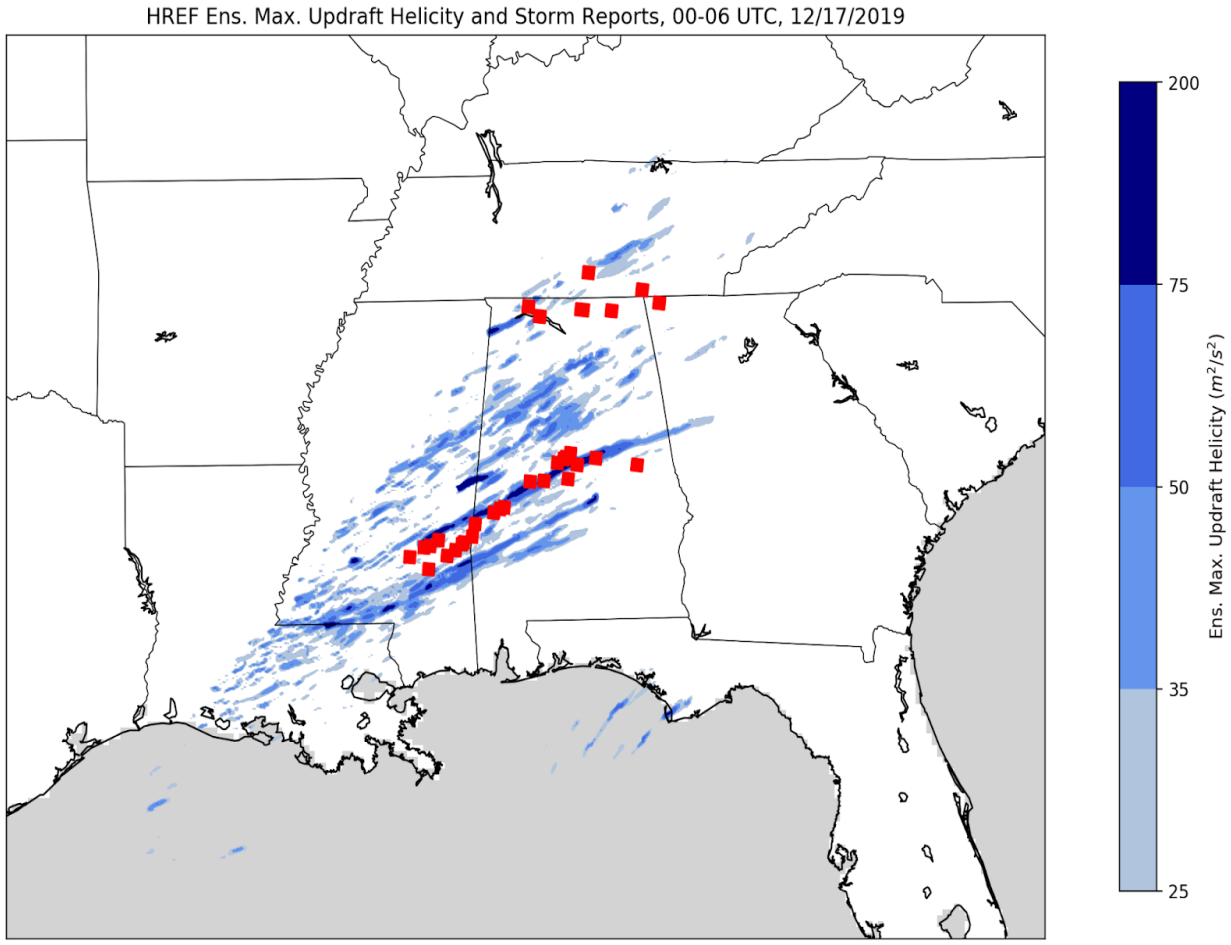

Figure 3.1. HREF 6-hr. Ensemble Maximum Updraft Helicity (blue shading) and Enhanced LSRs (red squares) plotted for 00-06 UTC 17 December 2019. Updraft Helicity thresholds are denoted by differing shades of blue, with deeper blues representing higher UH values. LSRs are enhanced using the 7x7 enhancement discussed in Section 2, which is what is used in verification.

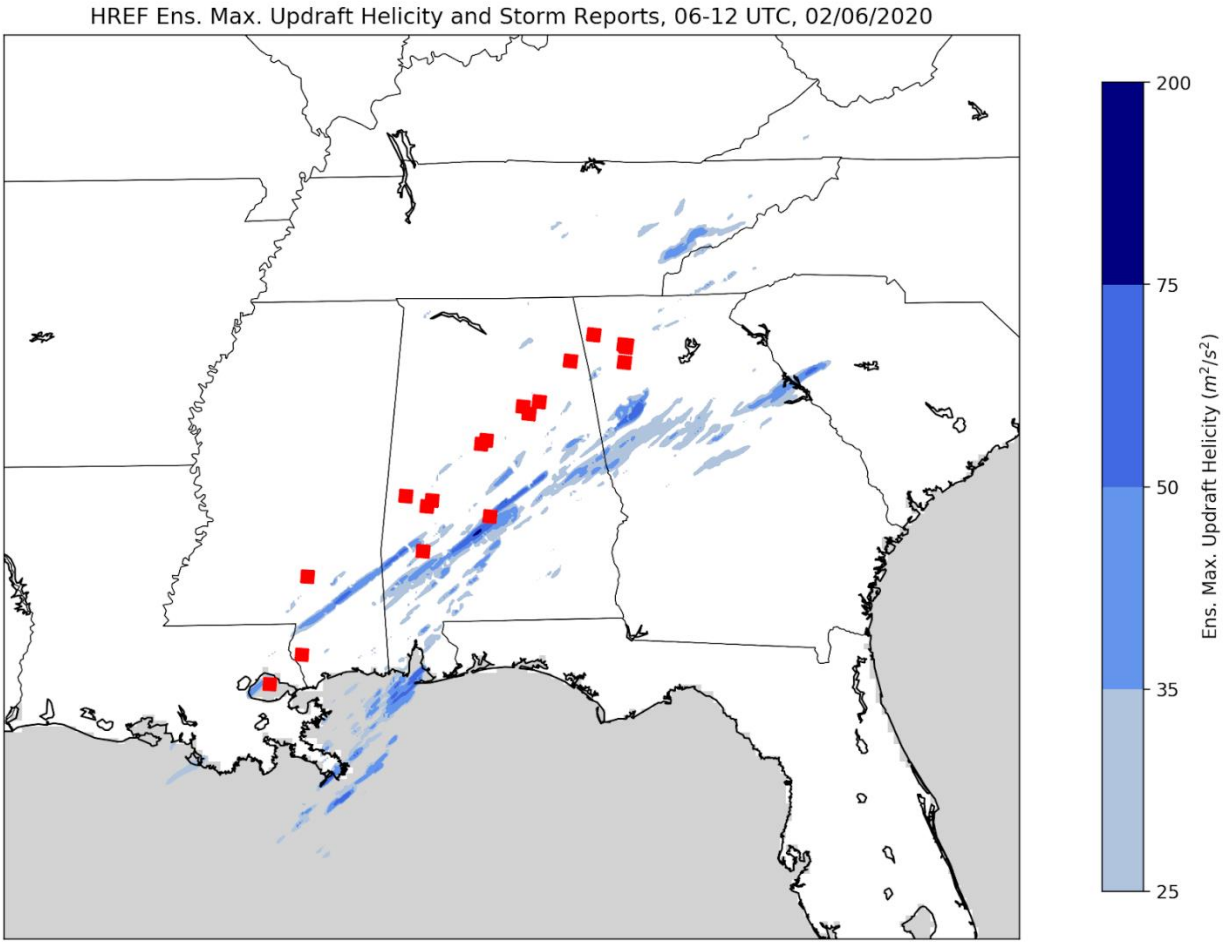

Figure 3.2. As in Figure 3.1, but for 06-12 UTC 6 February 2020.

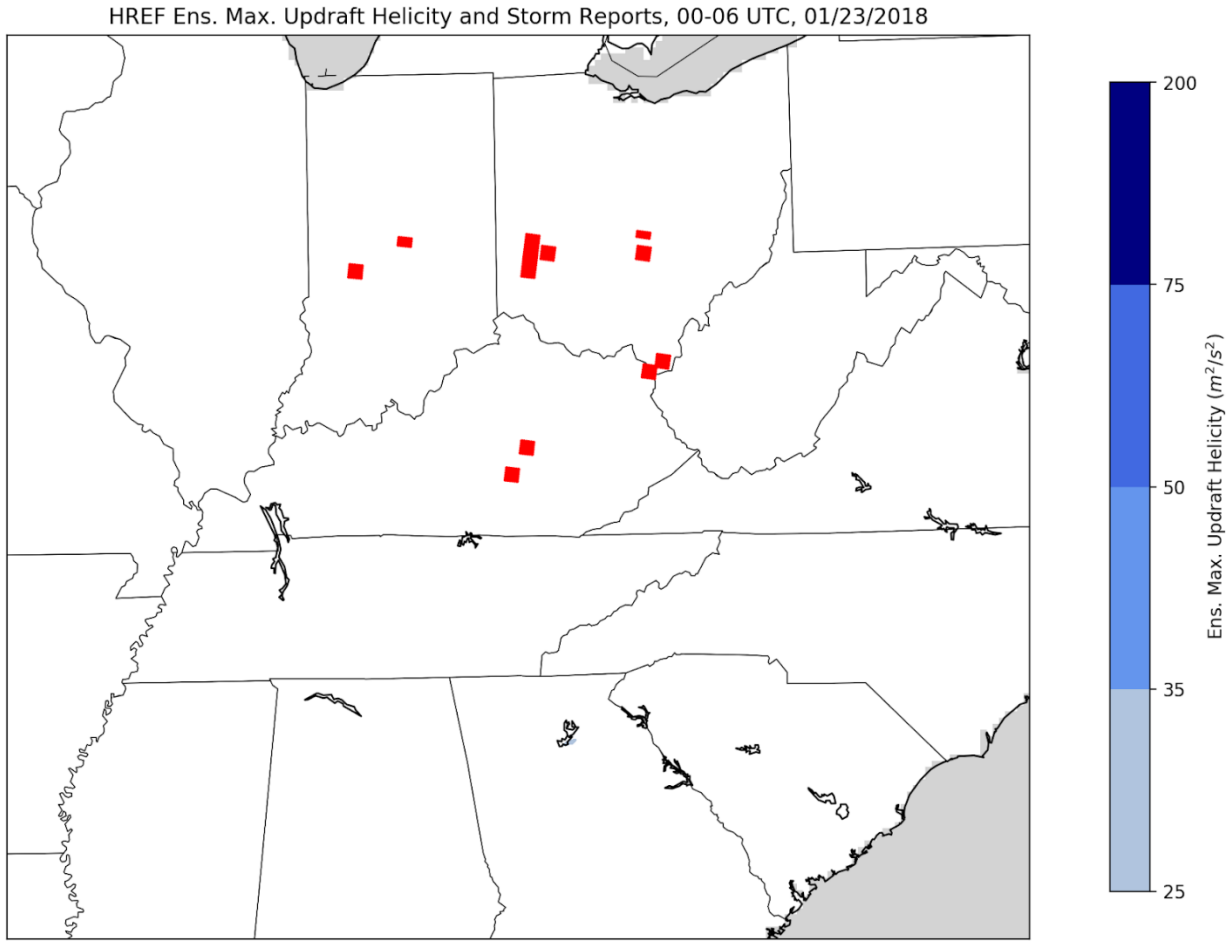

Figure 3.3. As in Figure 3.1, but for 00-06 UTC 23 January 2018.

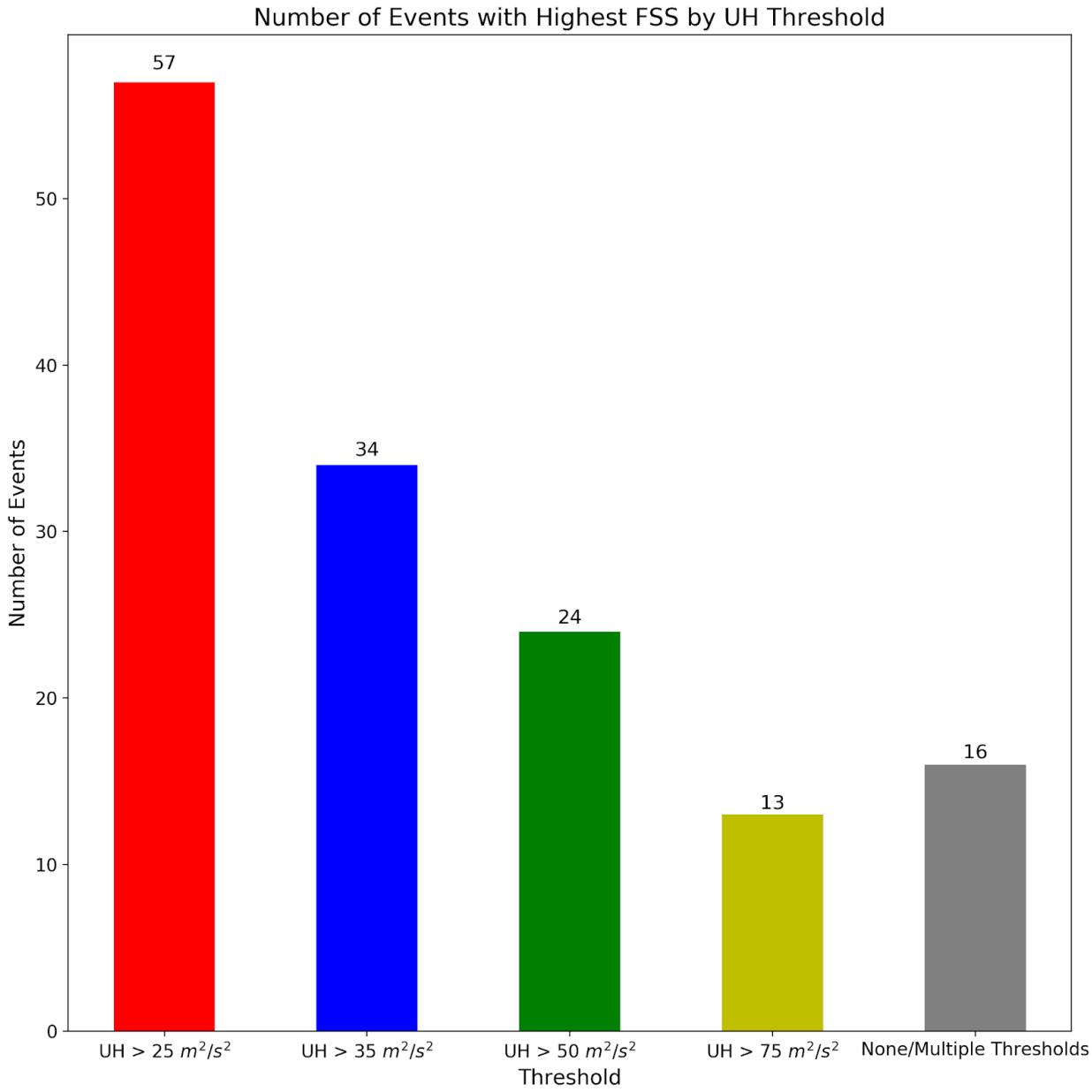

Figure 3.4. Number of events for which a specified threshold had the highest Fractions Skill Score (FSS) (Total Number of Events = 144). The "None/Multiple Thresholds" category includes events where either multiple thresholds had equally high FSS values or all thresholds had an FSS value of 0.

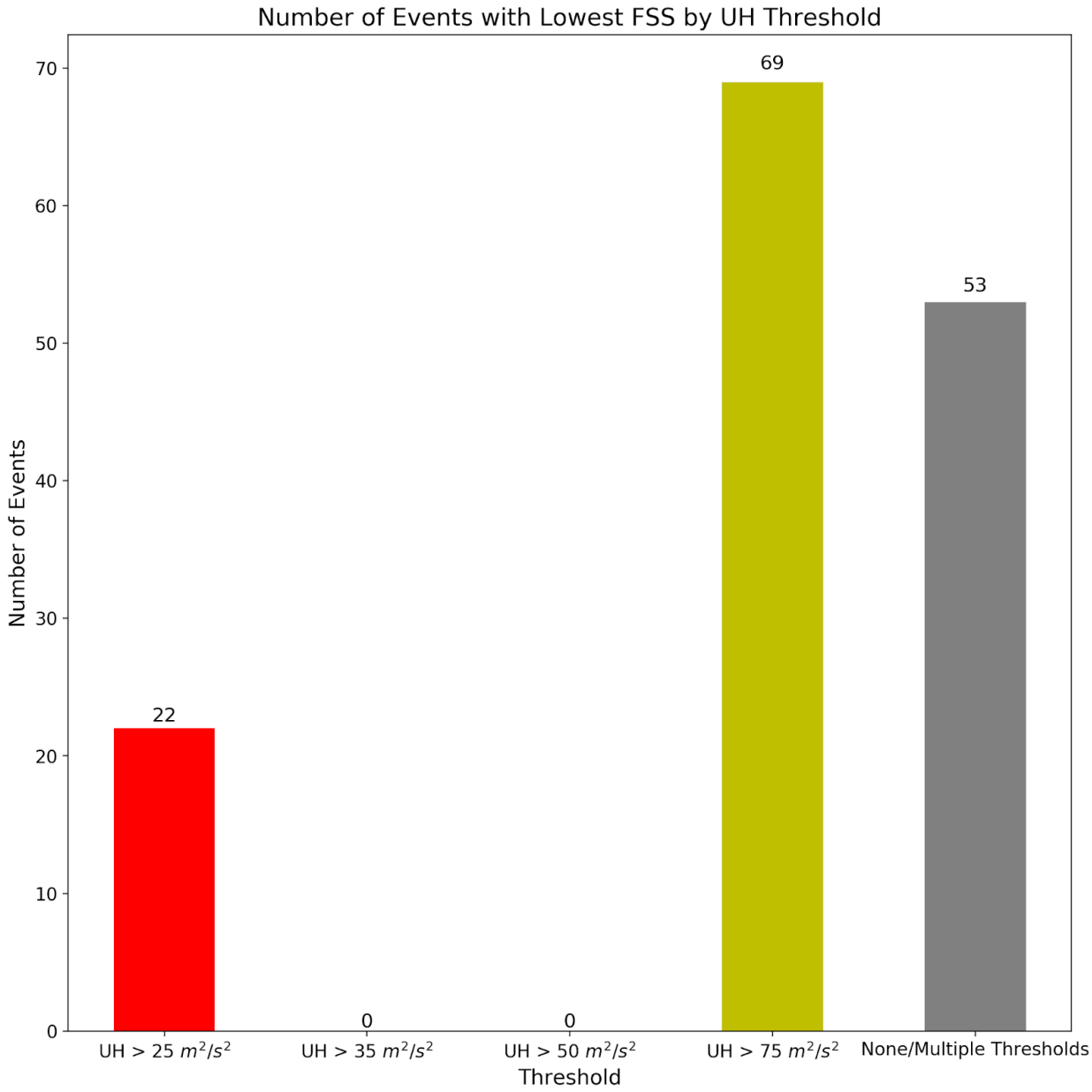

Figure 3.5. Number of events for which a specified threshold had the lowest Fractions Skill Score (FSS) (Total Number of Events = 144). The "None/Multiple Thresholds" category includes events where either multiple thresholds had equally low FSS values or all thresholds had an FSS value of 0.

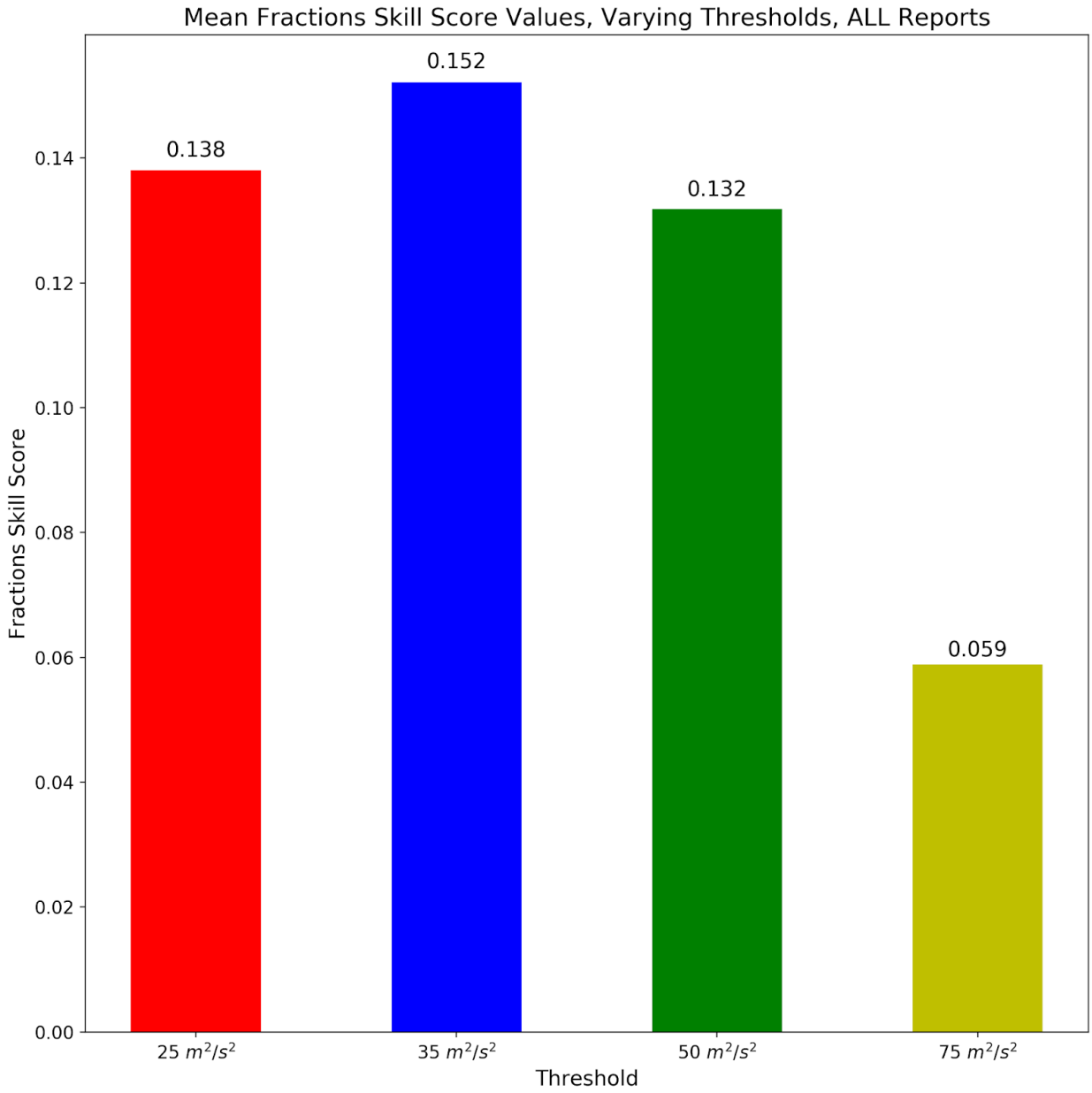

Figure 3.6. Mean FSS values by UH threshold over all events (n = 144), verifying against tornado and damaging wind storm reports. UH threshold increases from left to right, beginning with 25 m^2/s^2 (red), then 35 m^2/s^2 (blue), 50 m^2/s^2 (green), and finally 75 m^2/s^2 (yellow). The mean values for this plot were calculated weighting each event equally.

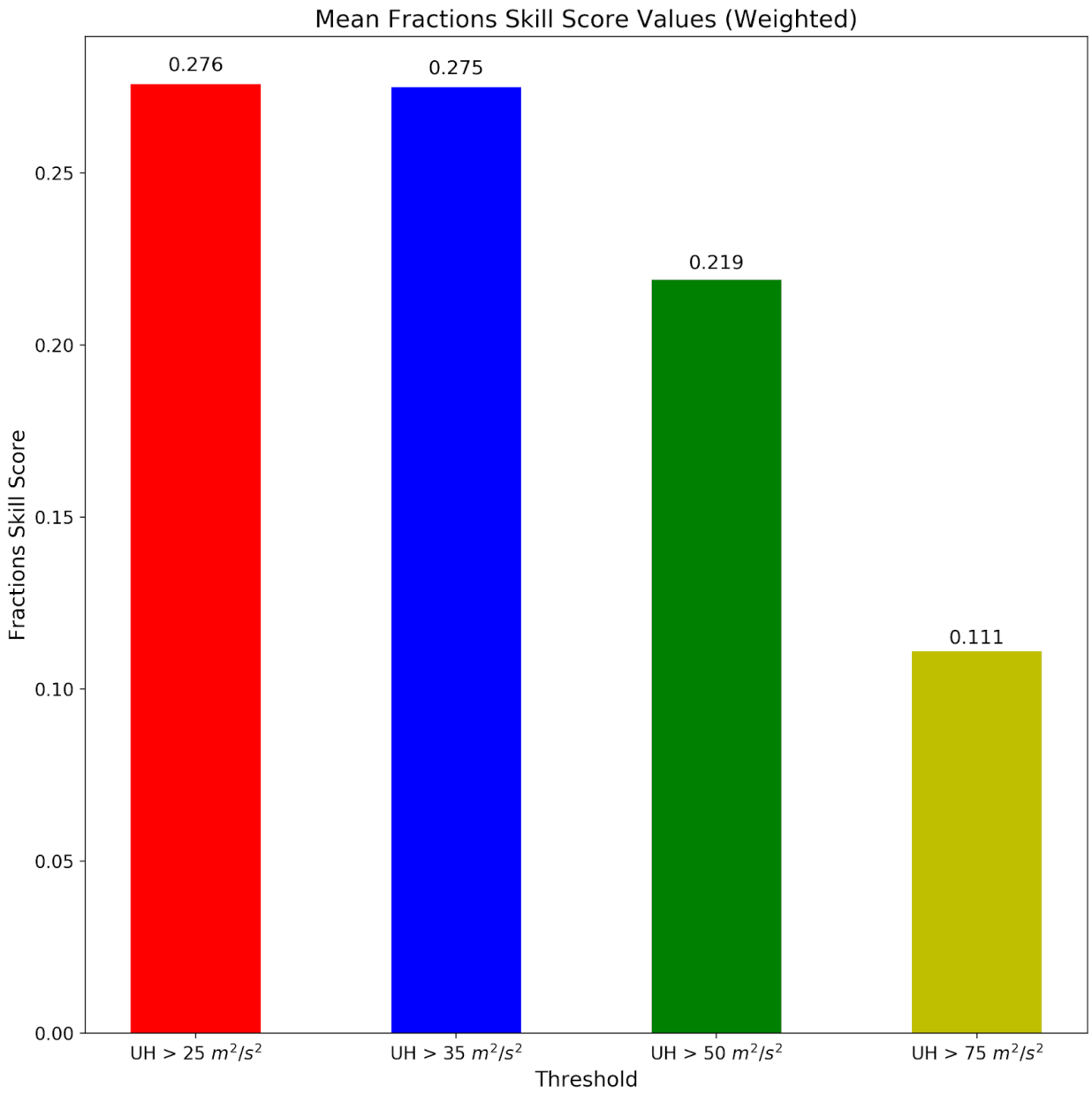

Figure 3.7. As in Figure 3.6, but mean FSS values were calculated by weighting each event depending on the total number of tornado and damaging wind reports that event had. See Section 3.1.2 and Equation 3.2 for more information on the weighting method.

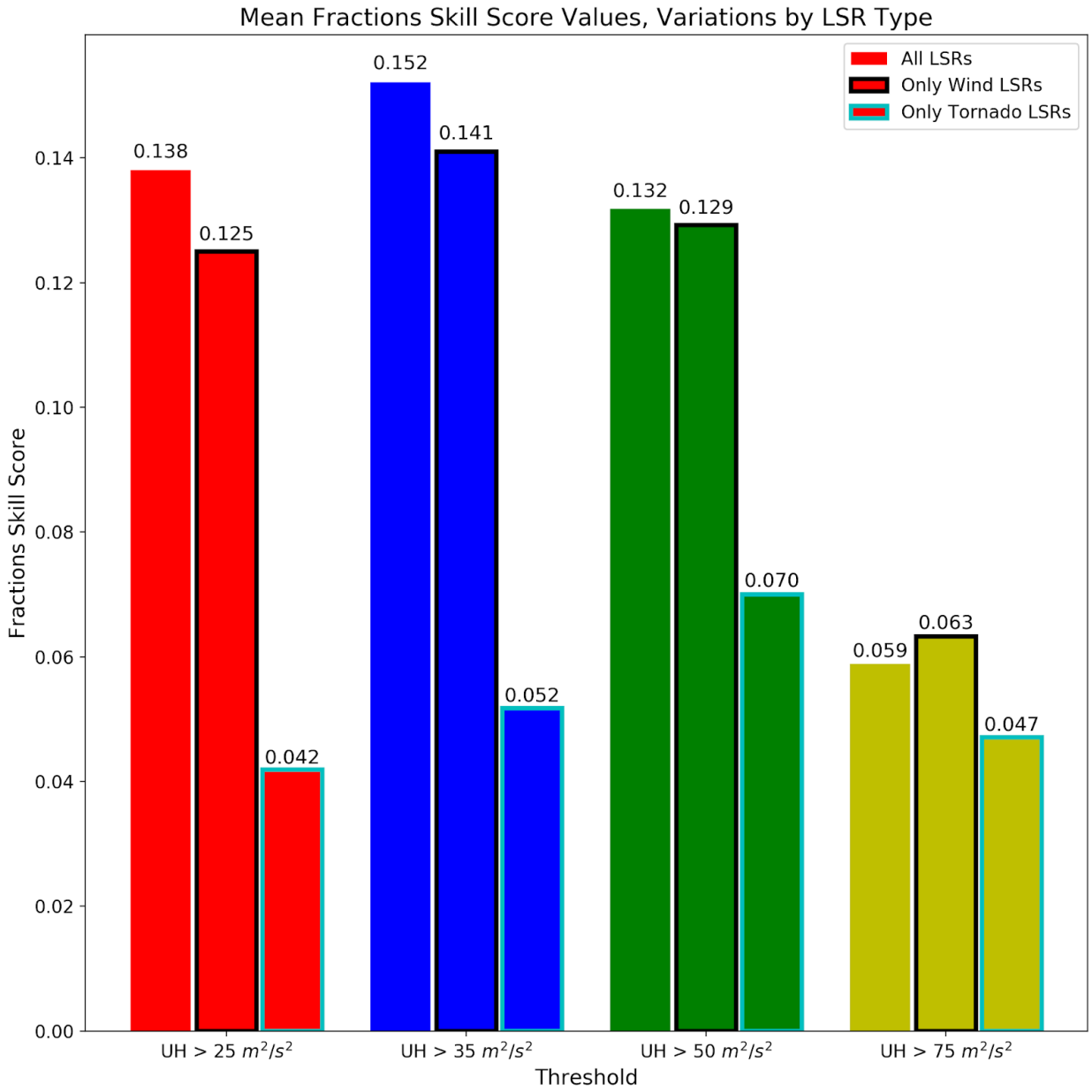

Figure 3.8. Mean FSS values by observation dataset over all events (n = 144). Verification against "All LSRs" (no outline) includes both tornado and damaging wind reports, with "Only Wind LSRs" (black outline) representing verification solely against damaging wind storm reports and "Only Tornado LSRs" (cyan outline) representing verification solely against tornado reports. The mean values for this plot were calculated weighting each event equally.

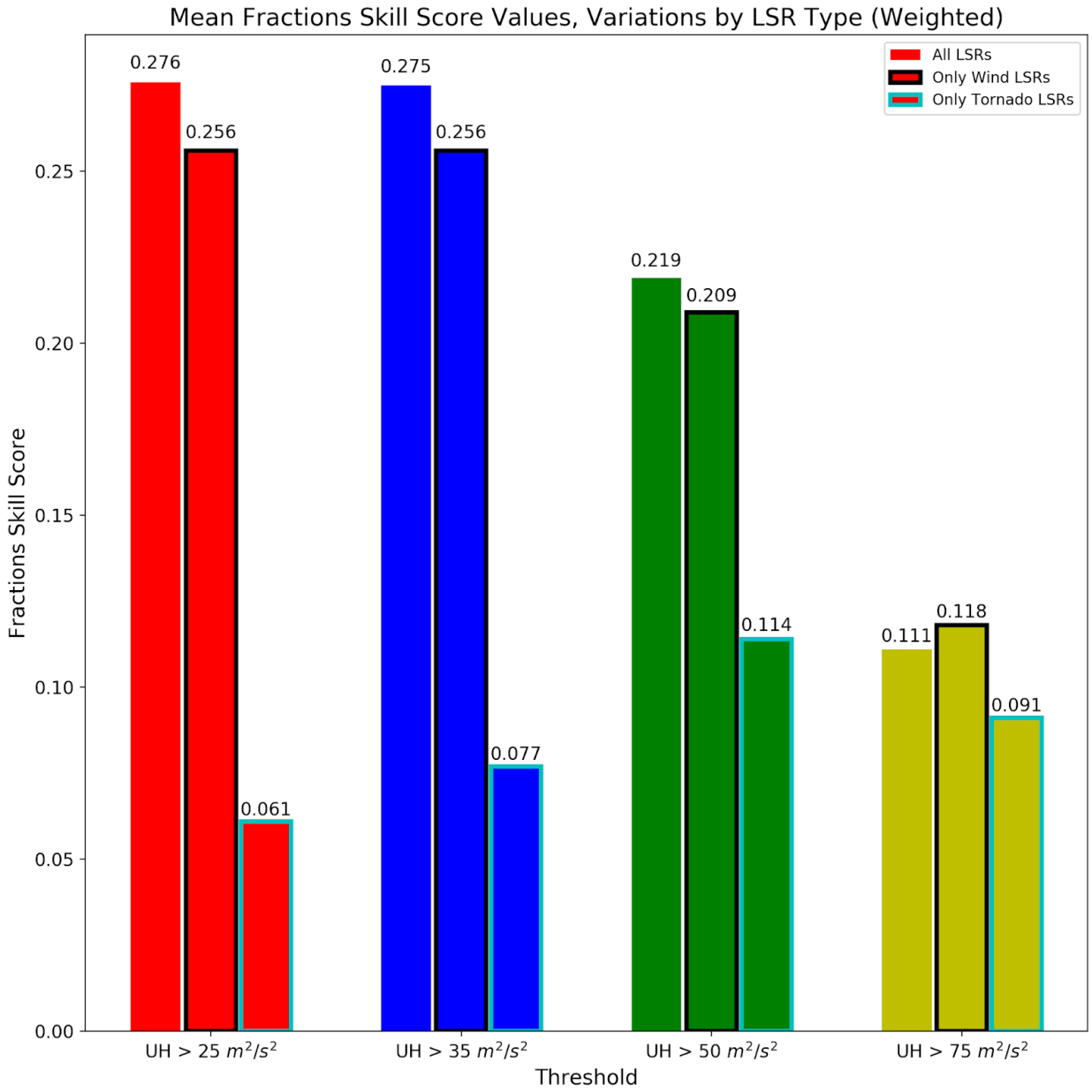

Figure 3.9. As in Figure 3.8, but mean FSS values were calculated by weighting each event depending on the total number of tornado and damaging wind reports that event had. See Section 3.1.2 and Equation 3.2 for more information on the weighting method.

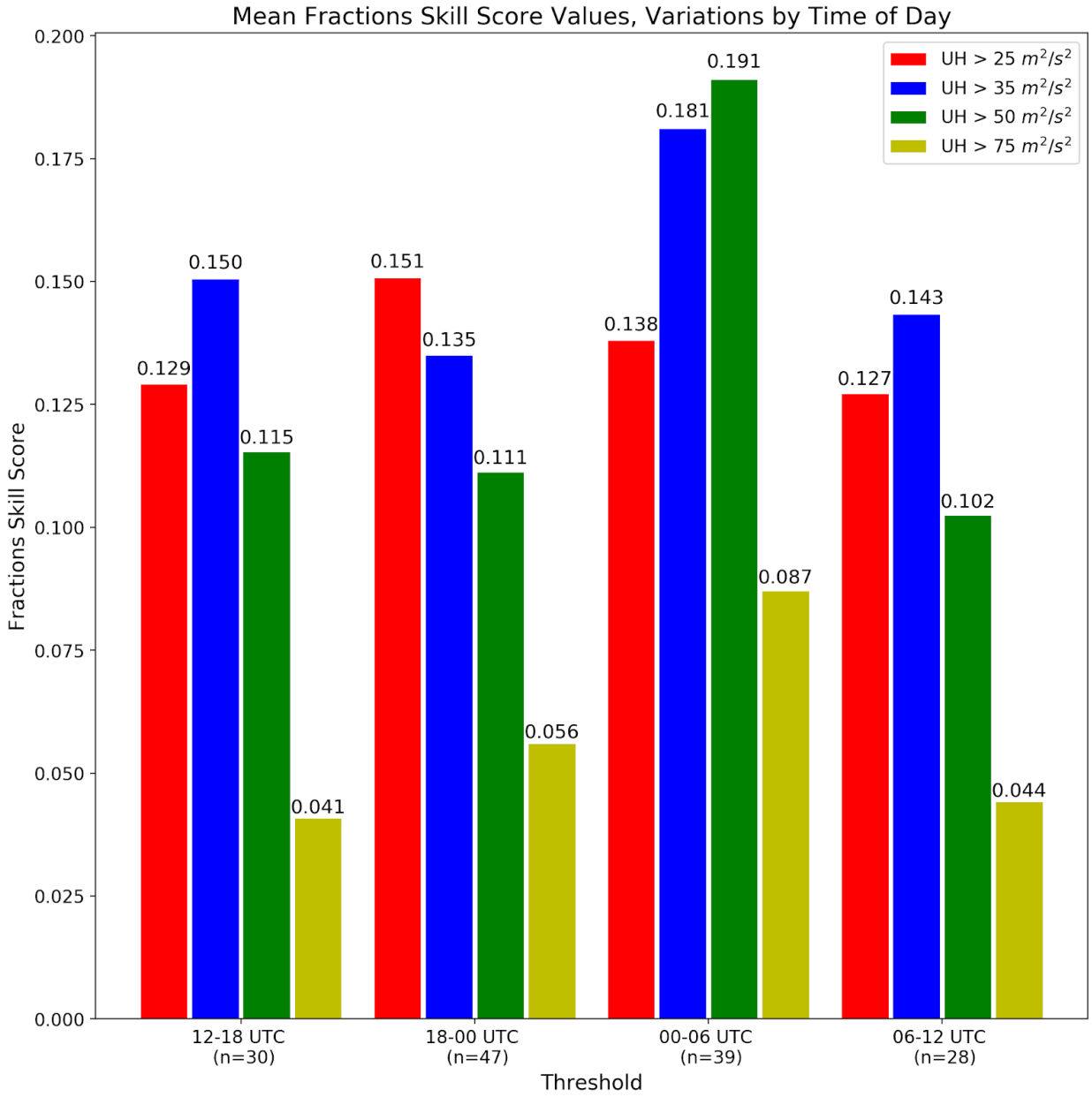

Figure 3.10. Mean FSS values, grouped by the six-hour window in which the event took place. The color scheme denoting which UH threshold is used is consistent with Figures 6-9. The leftmost grouping of bars represents the mean FSS for events which took place in the 12-18 UTC time frame, with groupings proceeding through the diurnal cycle from left to right. The mean values for this plot were calculated weighting each event equally.

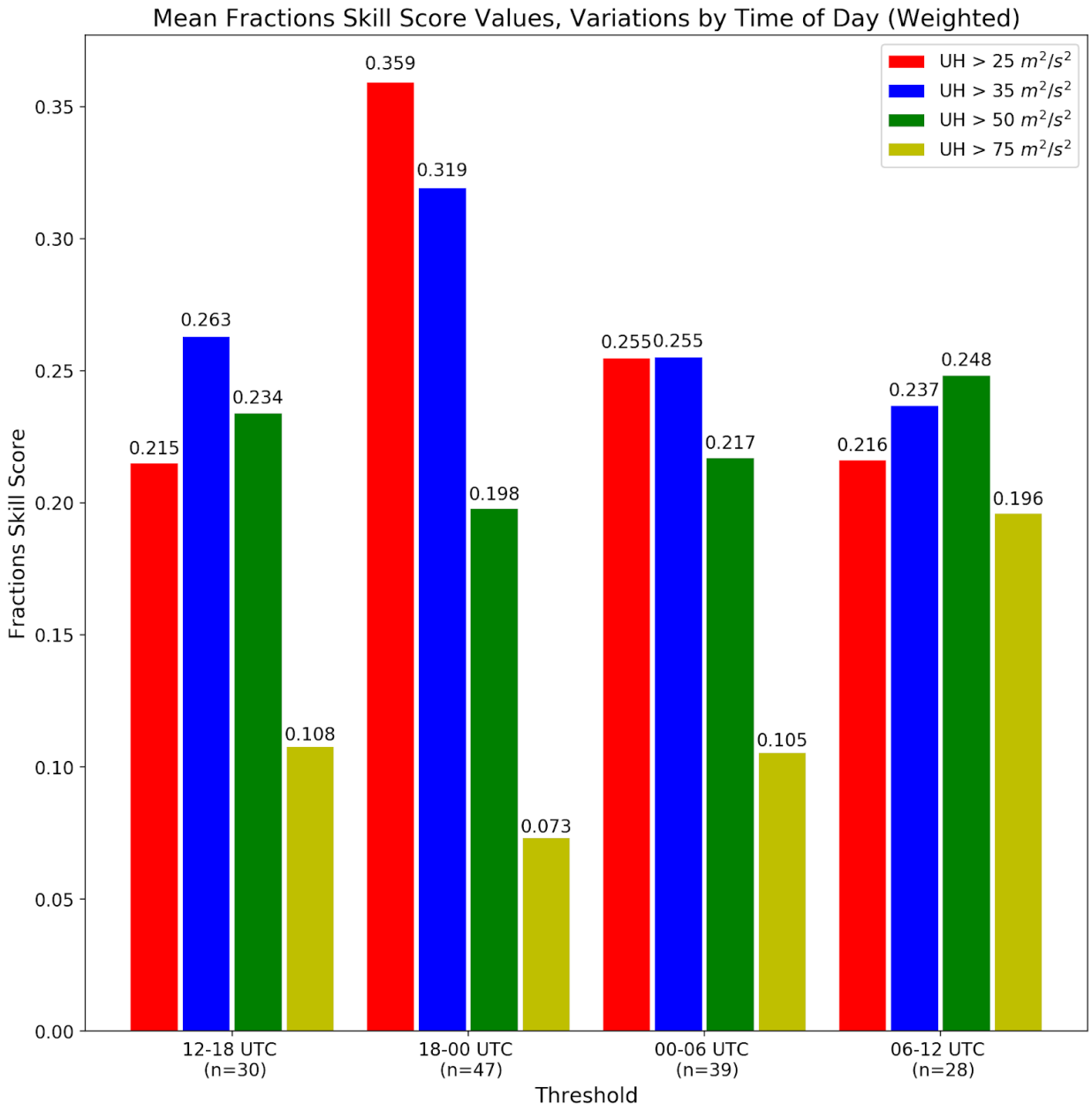

Figure 3.11. As in Figure 3.10, but mean FSS values were calculated by weighting each event depending on the total number of tornado and damaging wind reports that event had. See Section 3.1.2 and Equation 3.2 for more information on the weighting method.

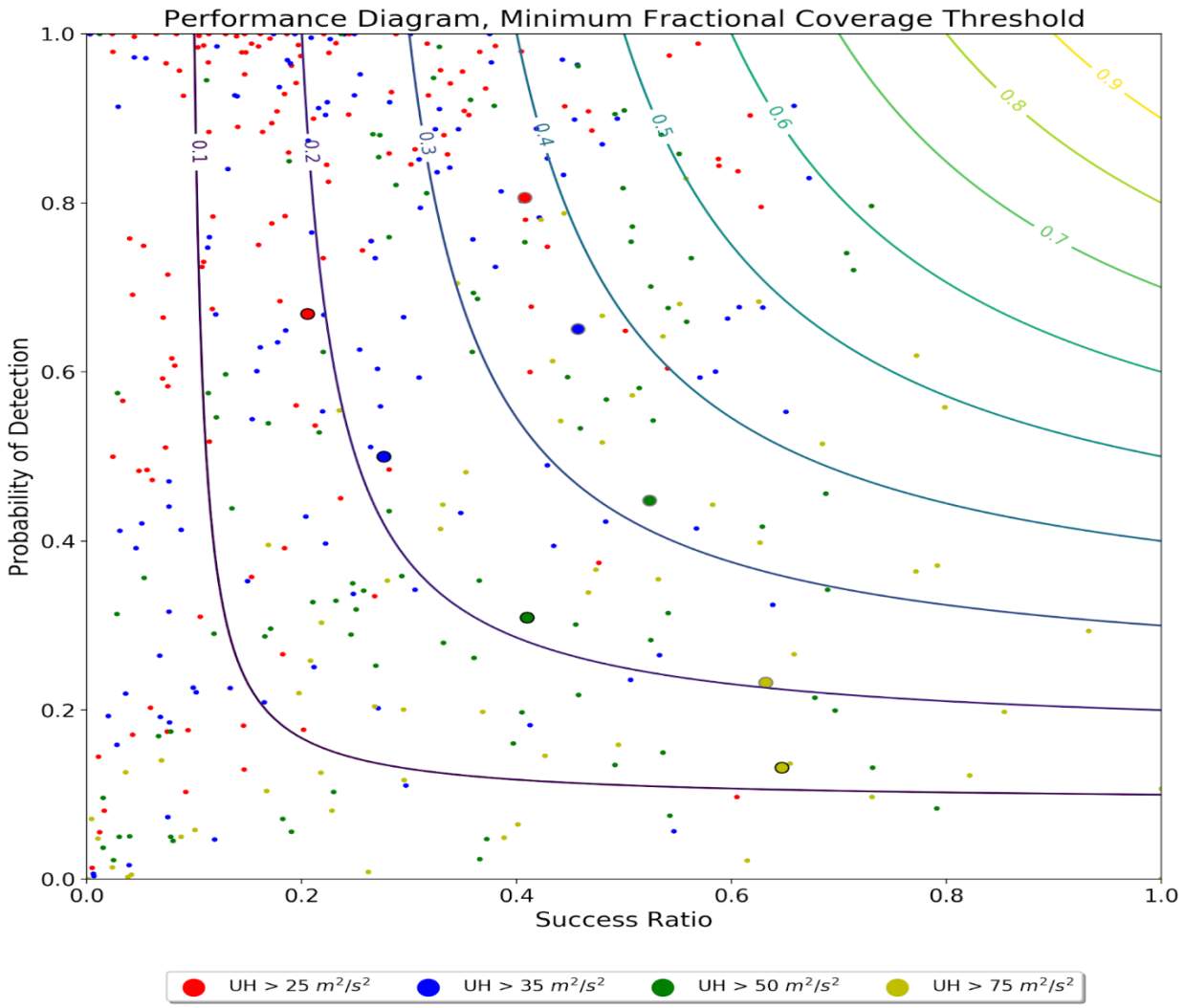

Figure 3.12. Performance Diagram using a minimum Fractional Coverage Threshold. Probability of Detection (POD) and Success Ratio (SR) ordered pairs are plotted for all events (*n* = 144) at the specified UH thresholds. Individual events are represented by the small dots, while the equally-weighted (black outline) and storm report-weighted (grey outline) mean POD and SR values are denoted by the large dots. The color scheme is based on UH threshold and is the same as in Figures 3.3-3.8. Lines of equal Critical Success Index (CSI) are indicated by the multicolored contours, starting from CSI = 0.1 and labeled at an interval of 0.1 up to CSI = 0.9.

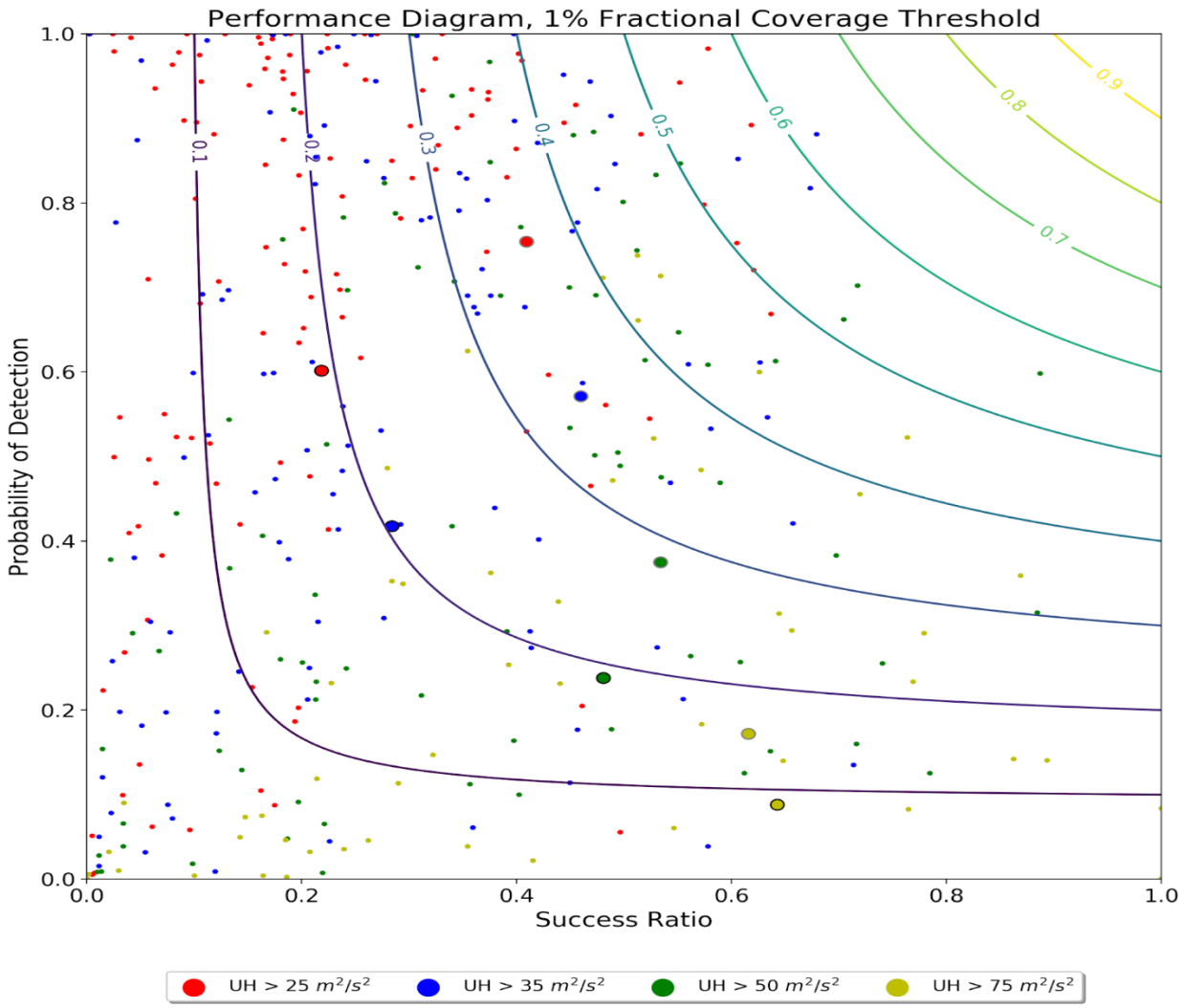

Figure 3.13. As in Figure 3.12, but for a 1% Fractional Coverage Threshold.

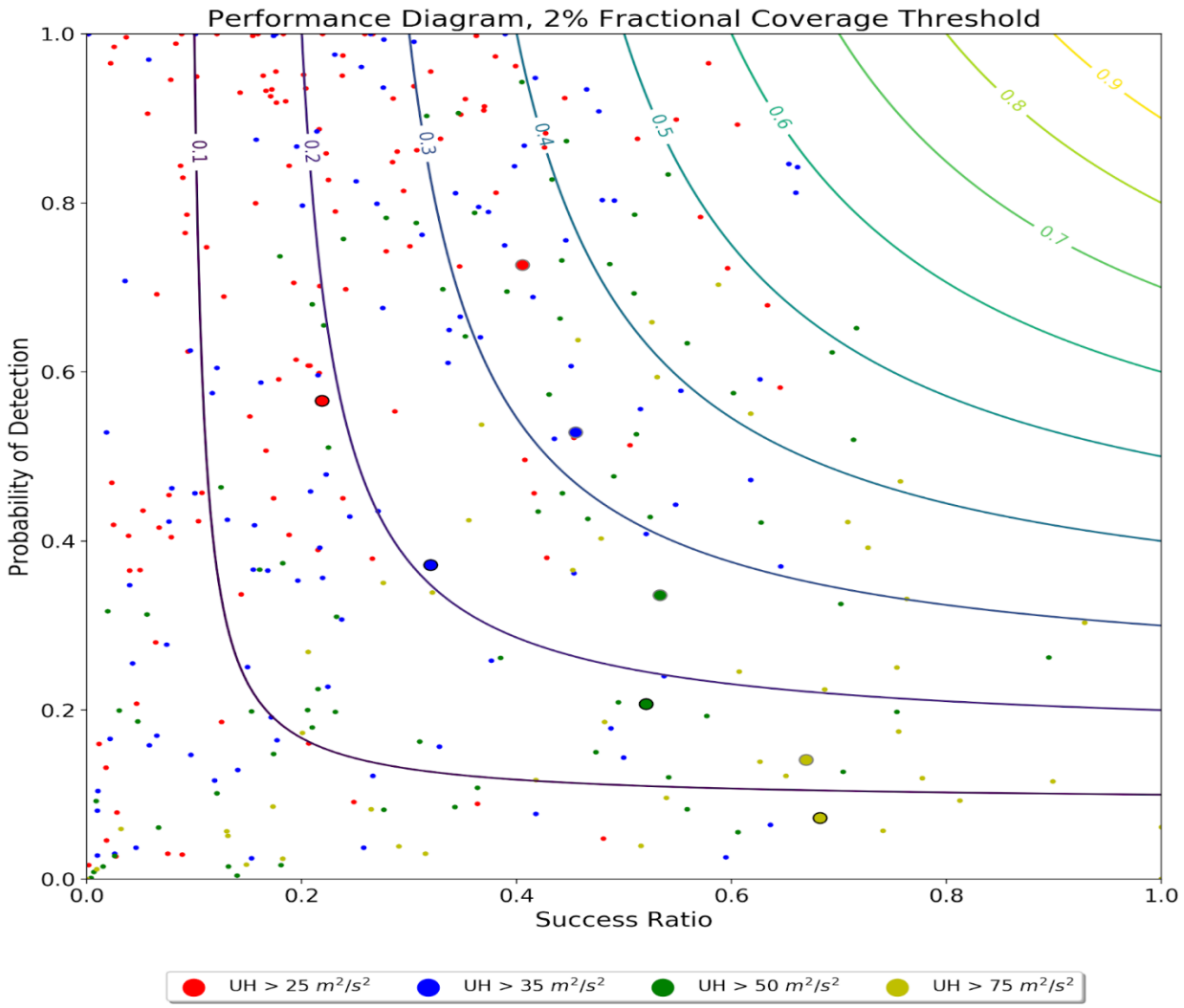

Figure 3.14. As in Figure 3.12, but for a 2% Fractional Coverage Threshold.

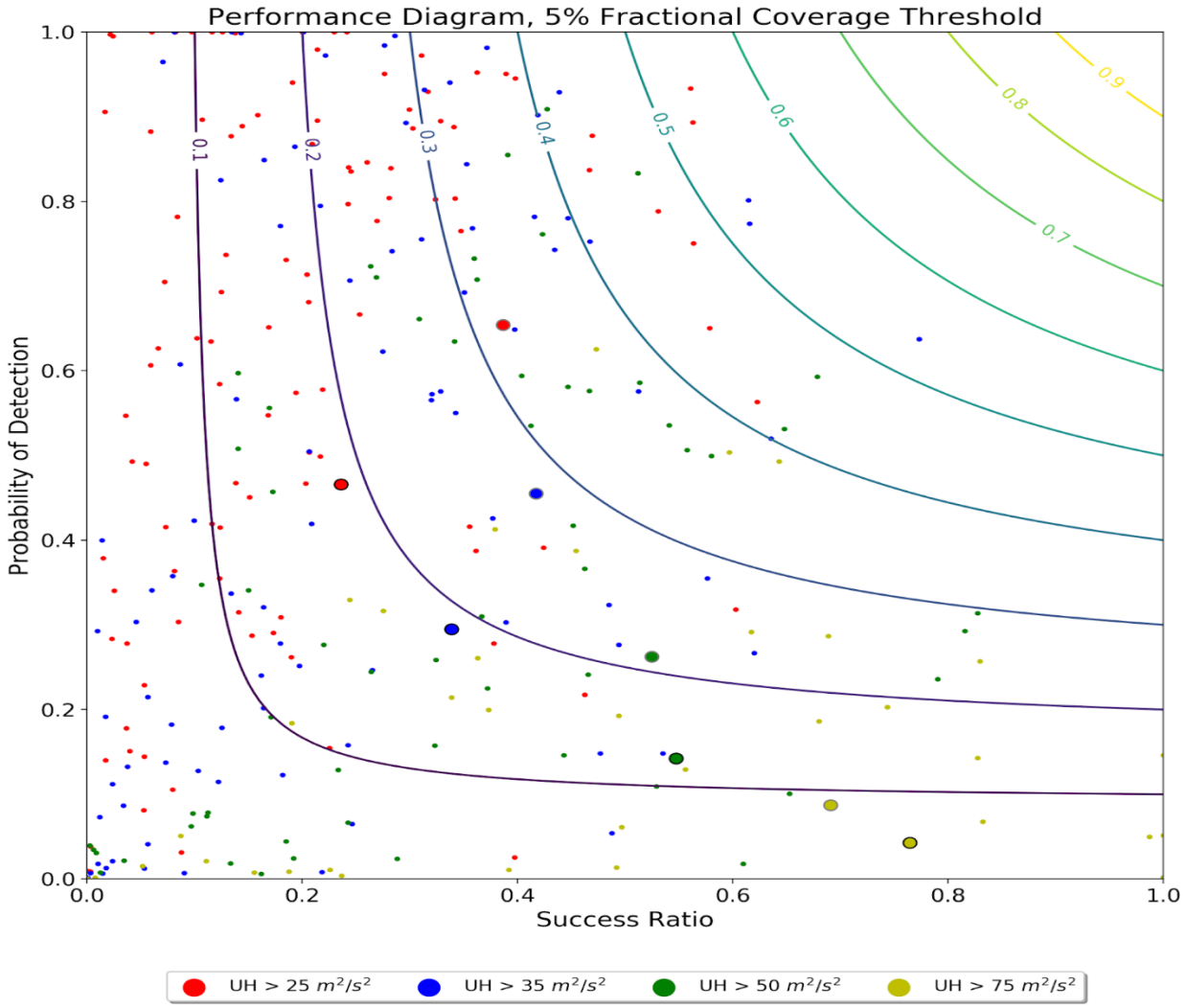

Figure 3.15. As in Figure 3.12, but for a 5% Fractional Coverage Threshold.

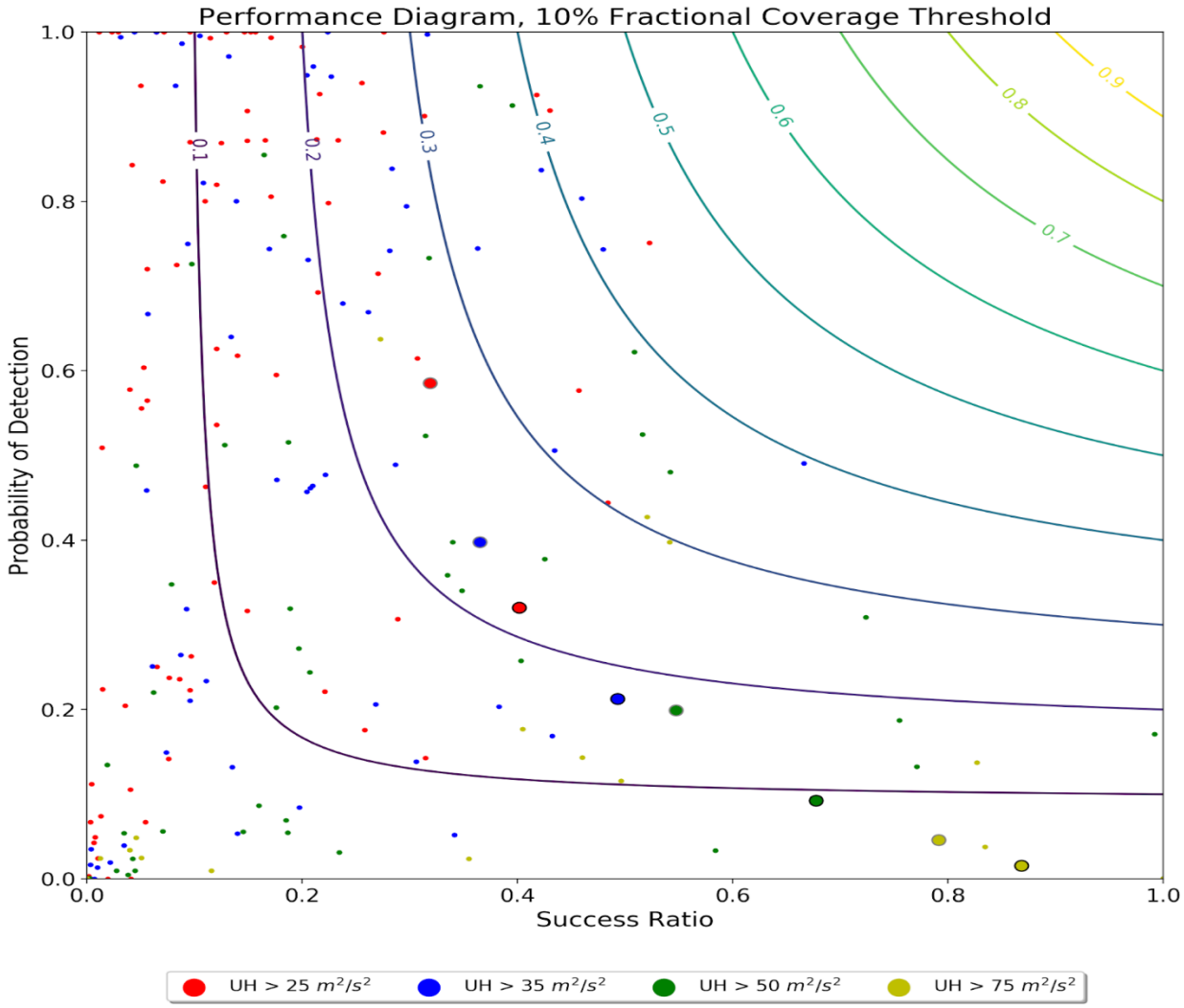

Figure 3.16. As in Figure 3.12, but for a 10% Fractional Coverage Threshold.

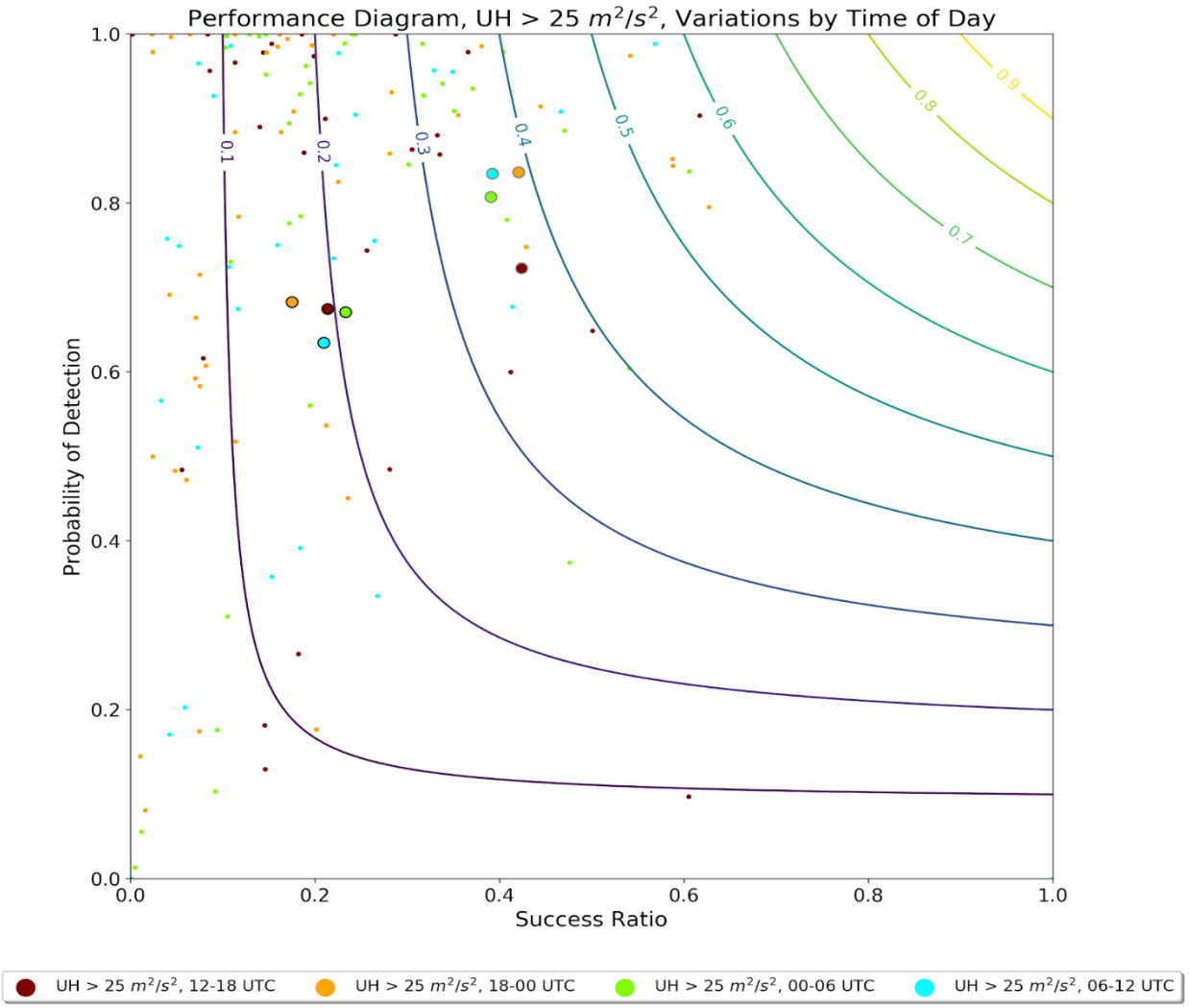

Figure 3.17. Performance Diagram using a minimum Fractional Coverage Threshold, highlighting variations in POD and SR values for events at different portions of the diurnal cycle. POD and SR ordered pairs are plotted for all events (n = 144) using the same UH threshold; in this plot, UH ≥ 25 m^2/s^2. Individual events are represented by the small dots, while the equally-weighted (black outline) and storm report-weighted (grey outline) mean POD and SR values are denoted by the large dots. The color scheme of the dots is based on the six-hour window which an event falls into, including 12-18 UTC (maroon, n = 30), 18-00 UTC (orange, n = 47), 00-06 UTC (chartreuse, n = 39), 06-12 UTC (cyan, n = 28). CSI contours are plotted as in Figure 3.9.

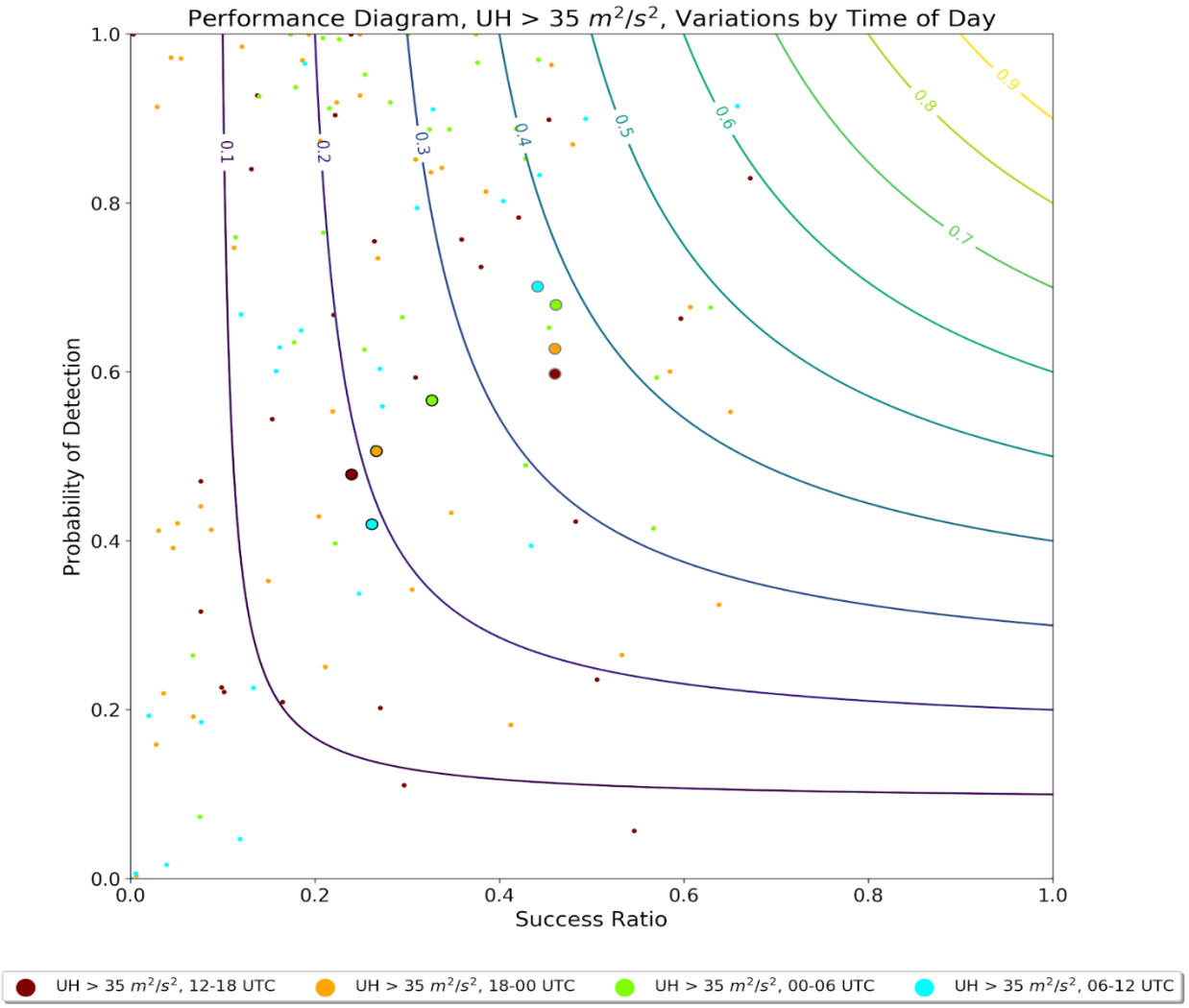

Figure 3.18. As in Figure 3.17, but for the threshold UH ≥ 35 m^2/s^2.

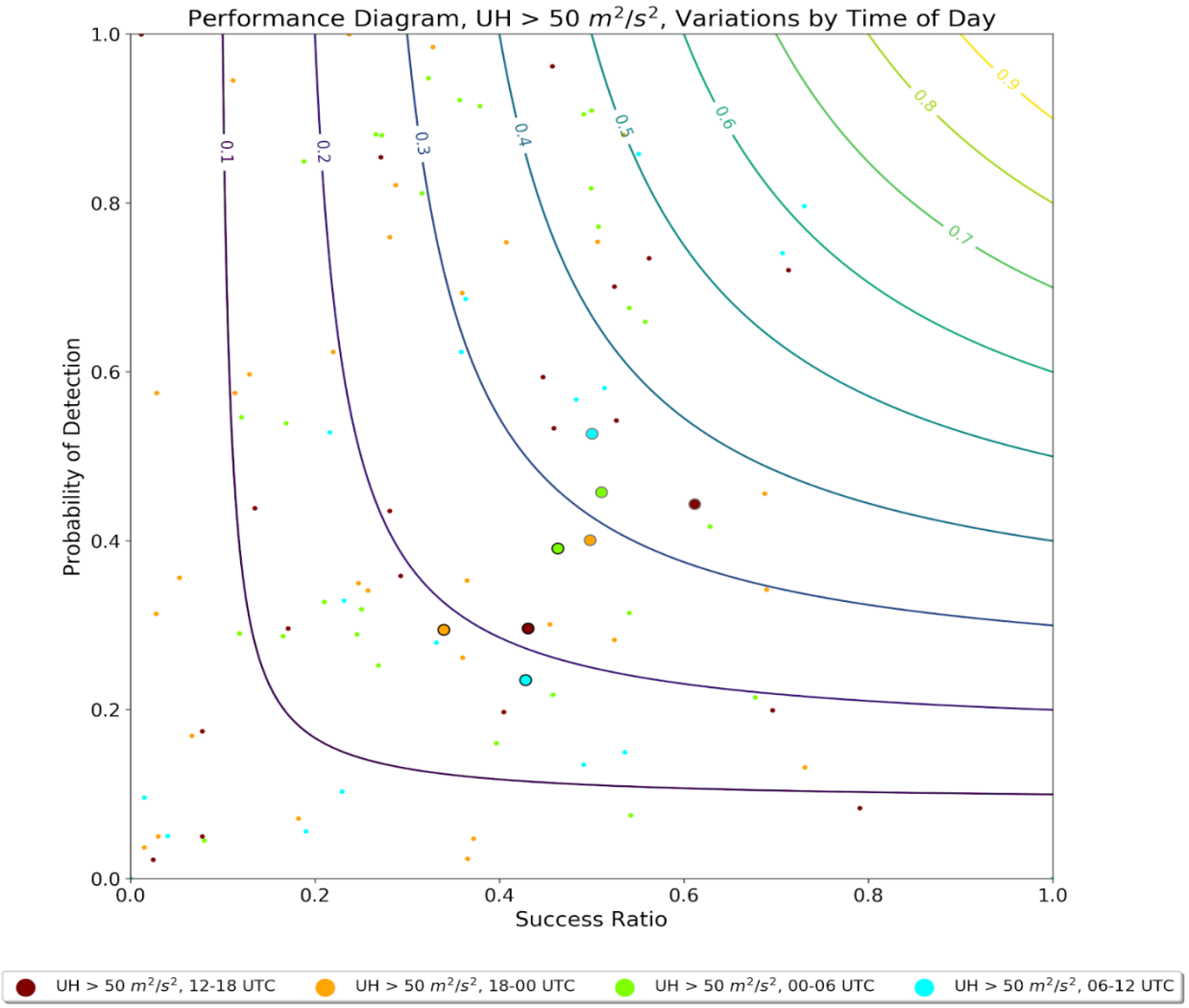

Figure 3.19. As in Figure 3.17, but for the threshold UH ≥ 50 m^2/s^2.

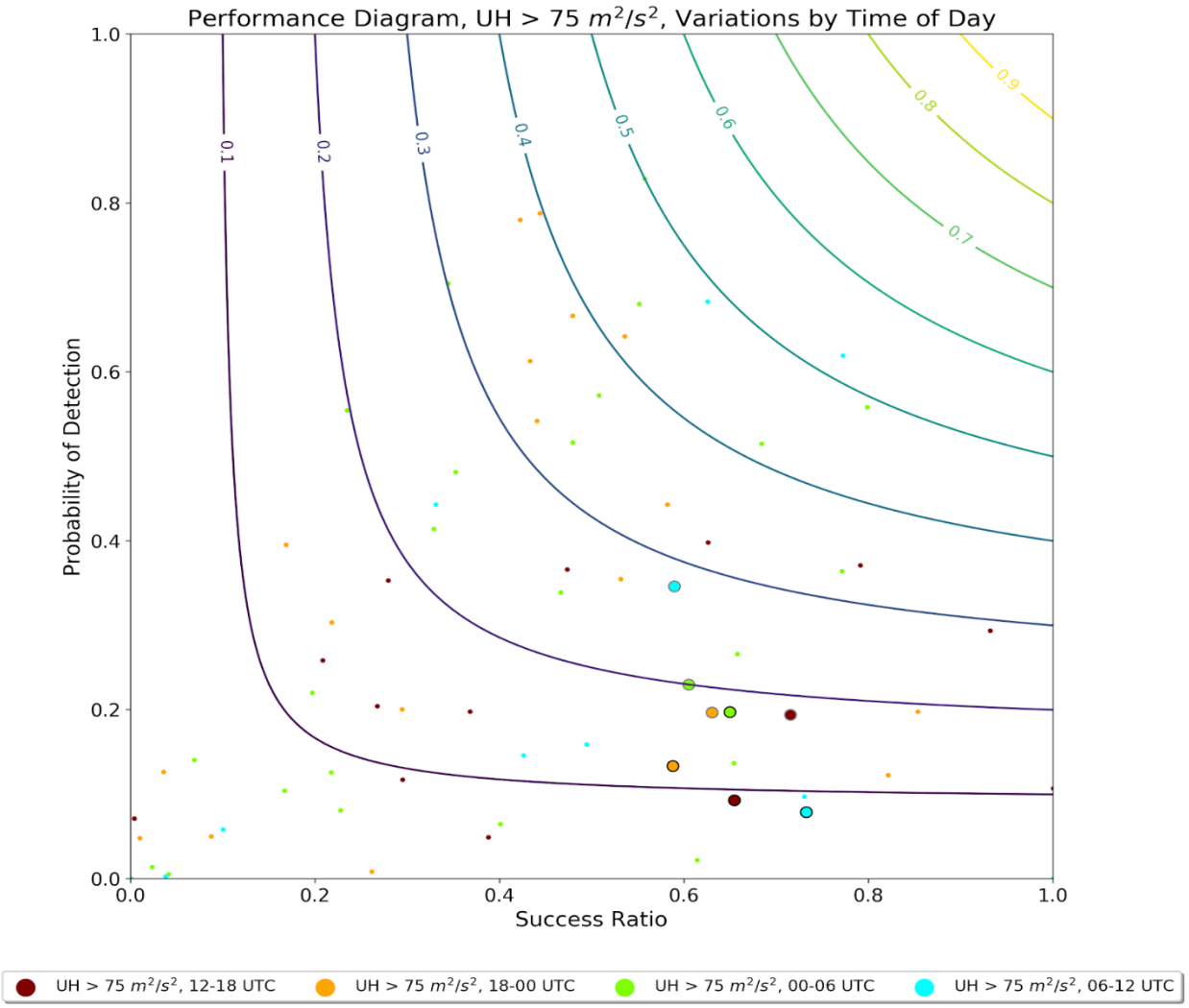

Figure 3.20. As in Figure 3.17, but for the threshold UH ≥ 75 m^2/s^2.

CHAPTER 4

Conclusions

Severe convection forecasting in HSLC environments is an arduous task at all forecast lead times, from nowcasting to day-ahead prediction. This is due, in part, to the fact that these storms frequently occur on small spatial and temporal scales that are difficult for traditional forecasting and remote sensing methods to predict and observe. Taking these things into consideration, it is vital that forecasters have as many tools as possible to try and predict these severe threats. Previous work by Sherburn and Parker (2014) and Sherburn et al. (2016) demonstrated that environmental-scale fields can be used to successfully predict severe hazards in HSLC environments. However, with the advent of CAM guidance, there is now the potential to derive additional information from explicit predictions of rotating updrafts or other severe convection proxies. To this point, very little work has been done to assess the skill of these explicit predictions from CAMs and CAM Ensembles, such as the HREF, in HSLC environments. That is what this thesis attempts to do.

4.1 Summary of Findings

In this study, HREF Ensemble Maximum 0-3 km UH forecasts were verified against LSRs for 144 six-hour events in the Southeastern United States, spanning over three cool seasons, 2017-2020. Neighborhood verification metrics, such as the Fractions Skill Score, and contingency table statistics were used in verification, with UH thresholds being set at 25, 35, 50, and 75 m^2/s^2 in order to convert the continuous UH field into a field of binary probabilities. While hail LSRs were excluded from verification, tornado and damaging wind LSRs were used jointly and separately as observation datasets, allowing for comparison of forecast skill between the two LSR types. Skill scores were aggregated over the entire set of events, and mean values for skill scores were calculated using both an unweighted mean and a mean which weighted events by the total amount of LSRs. The main takeaways from this study are:

- When verifying against all LSRs, Fractions Skill Score results suggest that UH ≥ 35 m^2/s^2 was the most skillful UH threshold on average, with UH ≥ 25 and 50 m^2/s^2 slightly behind and UH ≥ 75 m^2/s^2 well behind.
- Using an LSR-weighted mean, UH ≥ 25 and 35 m^2/s^2 are equally the most skillful UH threshold on average, with UH ≥ 50 m^2/s^2 slightly behind and UH ≥ 75 m^2/s^2 well behind. All mean skill values improved substantially when using an LSR-weighted mean, suggesting that events with more LSRs were associated with better forecast performance.
- For both the LSR-weighted and unweighted means, UH forecasts were more skillful when verifying against only wind reports than only tornado reports, although the gap in skill was smaller for UH ≥ 75 m^2/s^2.
- No consistent patterns were found in UH forecast performance across different portions of the diurnal cycle for either LSR-weighted or unweighted mean values.
- Unweighted mean contingency table metrics suggest that UH ≥ 35 and 50 m^2/s^2 are equally the most skillful UH thresholds, with UH ≥ 25 m^2/s^2 slightly behind and UH ≥ 75 m^2/s^2 well behind. Mean POD values monotonically decrease as UH threshold increases, while mean SR values consistently increase as UH threshold increases.
- LSR-weighted mean contingency table metrics suggest that UH ≥ 25 and 35 m^2/s^2 are equally most skillful, with UH ≥ 50 m^2/s^2 being slightly less skillful and UH ≥ 75 m^2/s^2 being substantially less skillful. POD values increase for all UH thresholds compared to unweighted mean values, while SR values increase for the lowest three threshold values.
- Despite noteworthy variations in skill between different UH thresholds, both Fractions Skill Score values and contingency table metrics suggest that all thresholds perform generally poorly on average, although some weakness in skill may be attributed to the methods of this study.

Ultimately, the most consistent finding across all of those mentioned above is that UH ≥ 75 m^2/s^2 is by far the least skillful UH threshold on average for Southeastern United States HSLC severe convection events. Given the findings of this study, it would be more reasonable for forecasters to use UH thresholds of below 75 m^2/s^2 when trying to anticipate HSLC severe convection. In addition, while the findings of this study suggest that 25 m^2/s^2 is broadly one of the most skillful thresholds, the extremely high amount of false alarms which occur when using

that threshold should cause forecasters to consider using a threshold that is somewhere in between that and 75 m^2/s^2. Outside of these broad conclusions, it is difficult to assess precisely which UH thresholds would be best for forecasters to use given the impact that our choice of LSR enhancement has on skill values including the perceived amount of overforecasting or underforecasting for each threshold.

4.2 Comparison to Prior UH Verification Studies

While there are certainly interesting patterns and relationships among the results presented in this study, it is important to frame these findings within the context of previous studies which conducted a verification of UH forecasts. While this study follows the methods of prior UH verification-related work, to the best of the author's knowledge, no antecedent studies have used the exact same study domain or forecast data as was outlined in Section 2; therefore, it is impossible to make an exact comparison between the findings of this study and findings of other studies. Nevertheless, the methodologies of prior studies discussed in this section are similar enough to this study to make useful comparisons between the findings of this study and those which preceded this study. Substantive differences between the study domain or data used of referenced antecedent studies and this study will be noted accordingly.

The mean FSS results suggest that, on average, HREF 0-3 km UH forecasts exhibit generally poor skill when using the verification methodology that is consistent with previous studies. Comparison to results from previous UH-based verification studies does not refute this assertion, although the mean skill values displayed in this study are certainly reasonable and competitive when placed in the context of other high-CAPE studies. Dawson et al. (2017) (herein D17) verified 2-5 km UH forecasts from an ensemble using the Advanced Research version of the WRF model (WRF-ARW) for 4 days with high-CAPE Central Plains severe convection in May 2013. For the neighborhood size which was most similar to the one used in this study (Circular Neighborhood with Radius of 50 km), FSS values varied between roughly 0.2 and 0.65 for the four severe weather days using the threshold UH $\geq$ 40 m^2/s^2 (their Figure 3). For UH $\geq$ 100 m^2/s^2, FSS values varied between roughly 0.1 and 0.5 for the four severe convection days (their Figure 4). While the mean unweighted FSS values from this study are either on the bottom fringe or outside the range of FSS values from D17, it should be noted that many individual events in this study had FSS values which fall into the range of FSS values in D17. Additionally,

the results of this study showed that mean FSS values increased substantially when events were weighted by number of LSRs, and all days examined in D17 had LSR counts which would be on the upper fringes of the distribution of LSR counts for events examined by this study. Therefore, it would be reasonable to compare to the D17 results using weighted mean FSS values, which compare more favorably, although weighted mean FSS values would still be toward the lower end of the range of FSS values from D17. While it would be difficult to make defensible generalizations about the overall performance of HREF 0-3 km UH forecasts in HSLC environments compared to the results of D17 given both the limited selection of events in D17 and the slightly indirect comparison between the two studies, it appears that FSS results from this study are reasonable in the context of a previous study which examined UH forecast skill in high-CAPE environments, albeit on the low side of the range of FSS values from the aforementioned study.

In addition to the FSS results, Sobash et al. (2011) (herein S11) provides a useful comparison for the contingency table results from this study. Like D17, S11 uses 2-5 km UH and verifies against LSRs; however, the S11 study uses deterministic output from the National Severe Storms Laboratory version of the WRF model (WRF-NSSL) during a study window which aligns with the 2008 NOAA Hazardous Weather Testbed Spring Experiment, 18 April - 8 June 2008. The spatial domain in S11 included the easternmost three-quarters of the conterminous United States. S11 calculated the same contingency table metrics as this study (POD, Bias, FAR/SR, CSI) on an 81-km grid, which is similar to the 75-km neighborhoods used for this study, for various UH thresholds between and including 33.75 m^2/s^2 and 103.25 m^2/s^2. The trends and patterns in contingency table statistics from S11 largely mirror those of this study. As UH threshold increases, bias, POD, FAR, and CSI all decrease (SR increases), which is a pattern also observed in the findings of this study (only considering thresholds from UH $\geq$ 35 m^2/s^2 upward to more closely align with the thresholds used by S11) (their Table 3). While no UH thresholds used in S11 precisely align with those of this study, some thresholds are close enough to allow for a reasonable comparison of results. For the low UH threshold (UH $\geq$ 35 m^2/s^2 in this study, UH $\geq$ 36 m^2/s^2 in S11), POD values are higher in this study (0.50 vs. 0.33), while SR values are higher in S11 (0.28 vs. 0.42), leading to similar CSI values (0.22 vs. 0.20). For the intermediate threshold (UH $\geq$ 50 m^2/s^2 in this study, UH $\geq$ 49.9375 m^2/s^2 in S11), POD values remain slightly higher for this study (0.31 vs. 0.25), while SR values remain slightly

higher in S11 (0.41 vs. 0.47), once again leading to similar CSI values (0.21 vs. 0.18). Finally, for the high UH threshold (UH ≥ 75 m^2/s^2 in this study, UH ≥ 77.5625 in S11), POD values are similar (0.13 vs. 0.14), SR values are higher in this study (0.65 vs. 0.52), and CSI values are similar (0.12 vs. 0.12). It is important to note that these comparisons are being made between HREF Ensemble Maximum UH and the WRF-NSSL deterministic UH output, so the fact that similar UH thresholds are being used and have roughly similar skill would suggest that useful UH thresholds will be lower in HSLC regimes than in high-CAPE regimes, which is what is primarily represented by S11. Nevertheless, the main point of comparing contingency table metrics is to show that patterns and trends in UH skill as well as overall skill values are roughly similar between this study and similar antecedent studies.

In this study, mean FSS values and contingency table statistics reveal that the lowest three UH thresholds used (UH ≥ 25, 35, and 50 m^2/s^2) are most skillful on average (with UH ≥ 35 m^2/s^2 slightly ahead of the other two thresholds), with UH ≥ 75 m^2/s^2 exhibiting lower amounts of skill on average. This is consistent with findings from Sobash and Kain (2017) (herein SK17), which looked at variations in most skillful 2-5 km UH thresholds by season for areas east of the Rocky Mountains in the conterminous United States. Using a similar model configuration to that in S11, SK17 showed that UH thresholds between 20-50 m^2/s^2 were associated with the highest average skill scores across most of the areas covered by the domain of this study during the cool season (their Figure 10a, b, and f). Further, SK17 calculated skill values for set UH thresholds of UH ≥ 20, 32, and 60 m^2/s^2, and many of the areas which exhibited the highest skill for the lower two thresholds in their study are located within the domain used for this study (their Figure 8a-c, d-f, and p-r). This is consistent with the findings of this study which suggest that 0-3 km UH thresholds of less than or equal to 50 m^2/s^2 are most skillful in the southeastern United States during the cool season.

Possibly the most direct comparison to this study in terms of model forecast data used comes from Roberts et al. 2020 (herein R20), which looked at variations in skill of 2-5 km UH forecasts over 21 days during the 2018 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (30 April - 1 June 2018) across the eastern two-thirds of the conterminous United States. R20 verified HREF v2.1 ensemble mean 2-5 km UH forecasts against LSRs and examined differences in forecast skill by UH threshold percentile (based on a UH climatology), finding that the greatest FSS values were achieved for ensemble mean UH thresholds between

the 85th and 90th percentile. This somewhat contradicts the results of this study since the UH thresholds with the highest skill corresponded to UH values between the 99th and 99.5th percentile ensemble maximum UH value. However, given that the R20 study did not discriminate between high-CAPE and low-CAPE environments, and that this study removed all UH values ≤ 0 m^2/s^2 when determining the UH climatology, it is certainly possible that differences in findings may arise simply from using different methodologies.

4.3 Future Work

There are several opportunities to expand upon and enhance the work started by this study, a few of which will be outlined in this section. First, in this study, although neighborhood verification methods allowed for some forgiveness in spatial differences between forecast and observation data, there was no smoothing applied to the forecast or observation data. The benefits of using spatial smoothing techniques are twofold: there is significant precedent in past studies with using spatial smoothing of UH forecast data (e.g., S11, SK17, R20), so using spatial smoothing would bring this work more into line with the antecedent literature, and spatial smoothing would create a forecast field which is more closely aligned with the purpose of high-resolution CAM guidance: to provide a short-term forecast of general locations where severe hazards are possible or likely. The aforementioned studies which used Gaussian smoothing all found that different amounts of spatial smoothing can greatly impact the skill scores, with some skill scores changing by as much as a factor of two with different amounts of smoothing.

While this study has tried to encapsulate HSLC severe convection as best as possible, one limitation of the current methods used was that the verification domain was fixed for all events. This creates the possibility (quite a likely one at that) that some areas within the verification domain exceeded the instability criterion set in Section 2 for some events. Ideally, in future studies, verification domain would be controlled by some instability parameter, possibly model output SBCAPE fields or gridded mesoanalysis of MUCAPE, in order to exclude verification of regions which do not fit into the HSLC paradigm. Additionally, because this study used 0-3 km UH forecasts, the NAM Nest members of the HREF had to be excluded in order to save time that would have been used calculating UH values from other available model output. In future studies, it may be beneficial to either add computed 0-3 km NAM Nest fields or simply use the more widely available 2-5 km UH products. This would make it easier to use ensemble

probabilities for verification instead of having to rely on the Ensemble Maximum or Ensemble Mean field.

Another limitation of the methods of this study was that HREF forecast data was only compiled on days with at least 1 LSR present within the study domain. While this was done primarily to save time and computational space, by only looking at days which had LSRs, this study misses out on capturing any events where UH forecasts exceeded the specified threshold, yet no severe reports were recorded. It is important to capture these "total false alarms" to provide a more holistic view of the performance of the HREF. While it would certainly be beneficial to have these additional null events, it is important to note that events with few LSRs can practically function like a false alarm event, and that 37 out of the 144 total events had only 1 or 2 LSRs. Finally, while the skill scores presented in this study were averaged across the entire database of HSLC severe convection events, it would be beneficial to examine the nuances and variance in performance from event to event. Furthermore, it would be worthwhile in future research to examine the environmental characteristics of individual events within this spatial and temporal domain to determine if there are any linkages between the synoptic- or mesoscale patterns of different well or poorly forecasted events.

Finally, while UH and LSRs are severe convection forecasts and observations with substantial precedent in verification research, other forecast and observation fields could be explored and performance and skill could be compared to that of this study. As was mentioned in Chapter 1, other severe convection proxies such as reflectivity, updraft strength, 10-meter maximum wind speed, or low-level vorticity could be used as alternative forecast proxies. Additionally, rotation track data could be used the observation dataset, with regions outside of 60 km from a NEXRAD site masked out of the verification domain. Ultimately, it is crucial that all of these severe proxies are analyzed to allow forecasters to have confidence that the CAM guidance tools they are using are effective at predicting severe convection hazards.

BIBLIOGRAPHY

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, https://doi.org/10.1175/WAF-D-16-0046.1.

Ashley, W. S., A. J. Krmenec, and R. Schwantes, 2008: Vulnerability due to nocturnal tornadoes. *Wea. Forecasting*, **23**, 795–807, https://doi.org/10.1175/2008WAF2222132.1.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.*, **131**, 2394–2416, https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2.

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, https://doi.org/10.1175/WAF-D-13-00127.1.

Dawson, L. C., G. S. Romine, R. J. Trapp, and M. E. Baldwin, 2017: Verifying supercellular rotation in a convection-permitting ensemble forecasting system with radar-derived rotation track data. *Wea. Forecasting*, **32**, 781–795, https://doi.org/10.1175/WAF-D-16-0121.1.

Dean, A. R., and R. S. Schneider, 2008: Forecast challenges at the NWS Storm Prediction Center relating to the frequency of favorable severe storm environments. Preprints, *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 9A.2. [Available online at https://ams.confex.com/ams/pdfpapers/141743.pdf.]

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Gensini, V. A., and M. K. Tippett, 2019: Global Ensemble Forecast System (GEFS) predictions of days 1-15 US tornado and hail frequencies. *Geophys. Res. Lett.*, **46**, 2922–2930, https://doi.org/10.1029/2018GL081724.

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

Janjić, Z., and R. L. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1 Dynamics. Tech. rep., NCAR Technical Note NCAR/TN-489+STR. https://opensky.ucar.edu:/islandora/object/technotes%3A502/.

Jensen, T., B. Brown, R. Bullock, T. Fowler, J. Halley Gotway, and K. Newman, 2020: Model Evaluation Tools Version 9.0.1 User's Guide. Tech. rep., Developmental Testbed Center, 482 pp, https://dtcenter.org/sites/default/files/community-code/met/docs/user-guide/MET_Users_Guide_v9.0.1.pdf.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC Storm-Scale Ensemble of Opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. Amer. Meteor. Soc., Nashville, TN, P9.137, https://www.spc.noaa.gov/publications/jirak/sseo_hwt.pdf.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley and Sons, 240 pp.

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

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

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, https://doi.org/10.1175/WAF-D-16-0141.1.

Potvin, C. K., C. Broyles, P. S. Skinner, H. E. Brooks, and E. Rasmussen, 2019: A Bayesian hierarchical modeling framework for correcting reporting bias in the U.S. tornado database. *Wea. Forecasting*, **34**, 15–30, https://doi.org/10.1175/WAF-D-18-0137.1.

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

——, and Coauthors, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea. Forecasting*, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

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

Roberts, N. M., 2005: An investigation of the ability of a storm scale configuration of the Met Office NWP model to predict flood-producing rainfall. Met Office Tech. Rep. 455, 80 pp.

——, and H. W. Lean, 2008: Scale-Selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Schneider, R. S., A. R. Dean, S. J. Weiss, and P. D. Bothwell, 2006: Analysis of estimated environments for 2004 and 2005 severe convective storm reports. Preprints, *23rd Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., 3.5. [Available online at https://ams.confex.com/ams/pdfpapers/115246.pdf.]

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

——, and Coauthors, 2015: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses. *Wea. Forecasting*, **30**, 1158–1181, https://doi.org/10.1175/WAF-D-15-0013.1.

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, https://doi.org/10.1175/WAF-D-13-00041.1.

——, ——, J. R. King, and G. M. Lackmann, 2016: Composite environments of severe and nonsevere high-shear, low-CAPE convective events. *Wea. Forecasting*, **31**, 1899–1927, https://doi.org/10.1175/WAF-D-16-0086.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF Version 3. Tech. rep., NCAR Technical Note NCAR/TN-475+STR. https://opensky.ucar.edu:/islandora/object/technotes%3A500/.

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

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

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

——, and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Wea. Forecasting*, **32**, 1885–1902, https://doi.org/10.1175/WAF-D-17-0043.1.

Strader, S. M., and W. S. Ashley, 2018: Finescale assessment of mobile home tornado vulnerability in the central and southeast United States. *Wea. Climate Soc.*, **10**, 797–812, https://doi.org/10.1175/WCAS-D-18-0060.1.

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, https://doi.org/10.1175/WAF-D-11-00116.1.

Wade, A. R., 2020: Dynamics and vorticity evolution in simulated low-CAPE supercells. Ph.D. dissertation, North Carolina State University, 94 pp, https://www.lib.ncsu.edu/resolver/1840.20/38196.

Yussouf, N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412, https://doi.org/10.1175/MWR-D-12-00237.1.

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

APPENDICES

Appendix A

Fractions Skill Scores for all 144 Events.

Event	Total LSRs	FSS, UH ≥ 25 m^2/s^2	FSS, UH ≥ 35 m^2/s^2	FSS, UH ≥ 50 m^2/s^2	FSS, UH ≥ 75 m^2/s^2
7 Nov 2017 00-06 UTC	2	0.017	0.006	0.001	0.000
7 Nov 2017 12-18 UTC	2	0.071	0.032	0.000	0.000
18 Nov 2017 18-00 UTC	94	0.351	0.033	0.001	0.000
19 Nov 2017 00-06 UTC	34	0.063	0.000	0.000	0.000
12 Jan 2018 18-00 UTC	2	0.000	0.000	0.000	0.000
13 Jan 2018 00-06 UTC	19	0.181	0.152	0.045	0.000
22 Jan 2018 00-06 UTC	7	0.309	0.463	0.334	0.000
22 Jan 2018 06-12 UTC	10	0.079	0.154	0.029	0.000
22 Jan 2018 12-18 UTC	5	0.117	0.186	0.000	0.000
23 Jan 2018 00-06 UTC	11	0.000	0.000	0.000	0.000
4 Feb 2018 12-18 UTC	2	0.010	0.083	0.205	0.000
7 Feb 2018 06-12 UTC	15	0.136	0.211	0.416	0.031
7 Feb 2018 12-18 UTC	16	0.169	0.380	0.106	0.000
7 Feb 2018 18-00 UTC	4	0.296	0.276	0.019	0.000
10 Feb 2018 18-00 UTC	3	0.001	0.000	0.000	0.000
11 Feb 2018 06-12 UTC	5	0.077	0.000	0.000	0.000
15 Feb 2018 18-00 UTC	1	0.148	0.237	0.447	0.000
16 Feb 2018 00-06 UTC	1	0.065	0.000	0.000	0.000
20 Feb 2018 18-00 UTC	1	0.040	0.062	0.036	0.000
21 Feb 2018 00-06 UTC	5	0.060	0.096	0.178	0.099
21 Feb 2018 06-12 UTC	2	0.090	0.145	0.000	0.000
24 Feb 2018 18-00 UTC	27	0.143	0.195	0.337	0.605
25 Feb 2018 00-06 UTC	70	0.242	0.324	0.453	0.384
25 Feb 2018 06-12 UTC	3	0.072	0.099	0.002	0.000
25 Feb 2018 18-00 UTC	1	0.000	0.000	0.000	0.000
28 Feb 2018 18-00 UTC	2	0.185	0.058	0.000	0.000
1 Mar 2018 06-12 UTC	2	0.028	0.000	0.000	0.000
1 Mar 2018 12-18 UTC	12	0.246	0.130	0.002	0.000
1 Mar 2018 18-00 UTC	21	0.091	0.043	0.008	0.000
17 Mar 2018 18-00 UTC	5	0.244	0.311	0.326	0.022
18 Mar 2018 00-06 UTC	11	0.062	0.088	0.049	0.001
27 Mar 2018 12-18 UTC	2	0.075	0.017	0.000	0.000
28 Mar 2018 18-00 UTC	19	0.231	0.314	0.499	0.530
29 Mar 2018 00-06 UTC	20	0.345	0.626	0.638	0.098
29 Mar 2018 06-12 UTC	8	0.191	0.090	0.007	0.000
29 Mar 2018 12-18 UTC	3	0.040	0.075	0.130	0.265
29 Mar 2018 18-00 UTC	2	0.005	0.006	0.000	0.000
3 April 2018 12-18 UTC	2	0.062	0.018	0.000	0.000
3 April 2018 18-00 UTC	86	0.291	0.423	0.460	0.086
4 April 2018 00-06 UTC	202	0.409	0.310	0.108	0.021
4 April 2018 06-12 UTC	14	0.077	0.001	0.000	0.000
7 April 2018 12-18 UTC	2	0.001	0.002	0.003	0.017
7 April 2018 18-00 UTC	5	0.012	0.009	0.000	0.000
31 Oct 2018 18-00 UTC	6	0.022	0.025	0.013	0.001
1 Nov 2018 00-06 UTC	33	0.188	0.342	0.540	0.208
1 Nov 2018 06-12 UTC	129	0.336	0.366	0.437	0.291
1 Nov 2018 12-18 UTC	17	0.108	0.190	0.283	0.010
1 Nov 2018 18-00 UTC	3	0.005	0.001	0.000	0.000
2 Nov 2018 18-00 UTC	9	0.035	0.044	0.004	0.000
3 Nov 2018 00-06 UTC	17	0.242	0.328	0.386	0.101
5 Nov 2018 12-18 UTC	4	0.000	0.000	0.000	0.000
5 Nov 2018 18-00 UTC	8	0.170	0.044	0.000	0.000
6 Nov 2018 00-06 UTC	45	0.203	0.390	0.630	0.278

6 Nov 2018 06-12 UTC	54	0.198	0.370	0.253	0.006
7 Nov 2018 18-00 UTC	10	0.068	0.025	0.004	0.000
1 Dec 2018 00-06 UTC	4	0.153	0.135	0.026	0.000
1 Dec 2018 06-12 UTC	69	0.256	0.067	0.002	0.000
20 Dec 2018 12-18 UTC	5	0.061	0.119	0.303	0.215
20 Dec 2018 18-00 UTC	4	0.004	0.007	0.014	0.032
21 Dec 2018 12-18 UTC	1	0.000	0.000	0.000	0.000
21 Dec 2018 18-00 UTC	7	0.004	0.000	0.000	0.000
27 Dec 2018 12-18 UTC	22	0.062	0.061	0.084	0.044
27 Dec 2018 18-00 UTC	2	0.003	0.002	0.001	0.001
31 Dec 2018 12-18 UTC	9	0.232	0.035	0.000	0.000
31 Dec 2018 18-00 UTC	47	0.565	0.343	0.101	0.000
4 Jan 2019 06-12 UTC	2	0.049	0.000	0.000	0.000
4 Jan 2019 18-00 UTC	4	0.009	0.000	0.000	0.000
5 Jan 2019 00-06 UTC	1	0.001	0.000	0.000	0.000
19 Jan 2019 06-12 UTC	8	0.277	0.181	0.023	0.000
19 Jan 2019 12-18 UTC	33	0.379	0.648	0.466	0.046
19 Jan 2019 18-00 UTC	13	0.121	0.273	0.491	0.220
20 Jan 2019 00-06 UTC	10	0.057	0.147	0.281	0.320
23 Jan 2019 12-18 UTC	1	0.044	0.184	0.084	0.000
23 Jan 2019 18-00 UTC	1	0.000	0.000	0.000	0.000
24 Jan 2019 00-06 UTC	11	0.139	0.261	0.484	0.020
24 Jan 2019 06-12 UTC	6	0.106	0.465	0.402	0.000
6 Feb 2019 06-12 UTC	2	0.016	0.011	0.000	0.000
6 Feb 2019 12-18 UTC	2	0.000	0.000	0.000	0.000
6 Feb 2019 18-00 UTC	2	0.005	0.007	0.006	0.000
7 Feb 2019 00-06 UTC	10	0.037	0.041	0.030	0.005
7 Feb 2019 06-12 UTC	5	0.058	0.106	0.215	0.399
7 Feb 2019 12-18 UTC	33	0.206	0.173	0.069	0.022
7 Feb 2019 18-00 UTC	2	0.013	0.000	0.000	0.000
8 Feb 2019 00-06 UTC	1	0.000	0.000	0.000	0.000
12 Feb 2019 06-12 UTC	2	0.003	0.000	0.000	0.000
12 Feb 2019 12-18 UTC	6	0.006	0.000	0.000	0.000
12 Feb 2019 18-00 UTC	51	0.388	0.039	0.000	0.000
13 Feb 2019 00-06 UTC	2	0.000	0.000	0.000	0.000
20 Feb 2019 00-06 UTC	2	0.000	0.000	0.000	0.000
23 Feb 2019 18-00 UTC	25	0.192	0.357	0.539	0.229
24 Feb 2019 00-06 UTC	7	0.094	0.156	0.265	0.384
24 Feb 2019 06-12 UTC	7	0.159	0.006	0.000	0.000
1 Mar 2019 12-18 UTC	2	0.074	0.012	0.000	0.000
1 Mar 2019 18-00 UTC	5	0.111	0.029	0.000	0.000
2 Mar 2019 00-06 UTC	1	0.000	0.000	0.000	0.000
2 Mar 2019 06-12 UTC	1	0.666	0.837	0.447	0.000
9 Mar 2019 12-18 UTC	28	0.399	0.592	0.383	0.026
9 Mar 2019 18-00 UTC	22	0.073	0.116	0.194	0.220
10 Mar 2019 00-06 UTC	21	0.108	0.184	0.318	0.345
14 Mar 2019 00-06 UTC	4	0.133	0.283	0.666	0.066
14 Mar 2019 06-12 UTC	7	0.000	0.000	0.000	0.000
14 Mar 2019 12-18 UTC	21	0.522	0.629	0.466	0.096
14 Mar 2019 18-00 UTC	93	0.546	0.311	0.126	0.029
15 Mar 2019 00-06 UTC	29	0.176	0.186	0.081	0.022
15 Mar 2019 18-00 UTC	1	0.001	0.000	0.000	0.000
16 Mar 2019 00-06 UTC	2	0.000	0.000	0.000	0.000
25 Mar 2019 18-00 UTC	12	0.392	0.040	0.001	0.000
26 Mar 2019 00-06 UTC	4	0.038	0.000	0.000	0.000
30 Mar 2019 18-00 UTC	8	0.365	0.459	0.086	0.000
31 Mar 2019 00-06 UTC	8	0.168	0.298	0.129	0.033
16 Dec 2019 12-18 UTC	16	0.123	0.112	0.096	0.048
16 Dec 2019 18-00 UTC	70	0.242	0.282	0.293	0.158

17 Dec 2019 00-06 UTC	35	0.225	0.398	0.686	0.396
17 Dec 2019 06-12 UTC	2	0.004	0.005	0.001	0.000
17 Dec 2019 12-18 UTC	16	0.252	0.145	0.001	0.000
17 Dec 2019 18-00 UTC	4	0.201	0.228	0.015	0.000
29 Dec 2019 00-06 UTC	14	0.142	0.286	0.107	0.000
29 Dec 2019 18-00 UTC	16	0.275	0.245	0.036	0.000
30 Dec 2019 00-06 UTC	4	0.022	0.019	0.000	0.000
30 Dec 2019 06-12 UTC	6	0.000	0.000	0.000	0.000
10 Jan 2020 18-00 UTC	6	0.084	0.109	0.150	0.256
11 Jan 2020 00-06 UTC	27	0.126	0.176	0.347	0.444
11 Jan 2020 06-12 UTC	142	0.212	0.263	0.367	0.484
11 Jan 2020 12-18 UTC	168	0.326	0.384	0.409	0.273
11 Jan 2020 18-00 UTC	247	0.611	0.645	0.275	0.044
12 Jan 2020 00-06 UTC	134	0.365	0.194	0.055	0.001
12 Jan 2020 06-12 UTC	4	0.116	0.095	0.000	0.000
5 Feb 2020 18-00 UTC	29	0.170	0.237	0.220	0.148
6 Feb 2020 00-06 UTC	32	0.188	0.138	0.047	0.018
6 Feb 2020 06-12 UTC	23	0.085	0.183	0.128	0.007
6 Feb 2020 12-18 UTC	52	0.181	0.253	0.340	0.161
6 Feb 2020 18-00 UTC	71	0.222	0.364	0.172	0.000
7 Feb 2020 00-06 UTC	49	0.123	0.115	0.064	0.003
7 Feb 2020 06-12 UTC	7	0.028	0.048	0.011	0.000
7 Feb 2020 12-18 UTC	87	0.013	0.004	0.000	0.000
12 Feb 2020 18-00 UTC	14	0.125	0.063	0.000	0.000
13 Feb 2020 00-06 UTC	28	0.352	0.411	0.008	0.000
13 Feb 2020 06-12 UTC	3	0.000	0.000	0.000	0.000
13 Feb 2020 12-18 UTC	12	0.019	0.000	0.000	0.000
13 Feb 2020 18-00 UTC	1	0.000	0.000	0.000	0.000
3 Mar 2020 00-06 UTC	19	0.337	0.495	0.484	0.142
3 Mar 2020 06-12 UTC	21	0.239	0.310	0.126	0.016
3 Mar 2020 12-18 UTC	5	0.074	0.048	0.027	0.000
3 Mar 2020 18-00 UTC	1	0.021	0.077	0.350	0.046
Totals	**3148**	-	-	-	-
Unweighted Mean	-	**0.138**	**0.152**	**0.132**	**0.059**
Weighted Mean	-	**0.276**	**0.275**	**0.219**	**0.111**

Appendix B

Mean Contingency Table Statistics, Variations by Fractional Coverage Threshold.

Quantity	Minimum FCT	1% FCT	2% FCT	5% FCT	10% FCT
$\overline{POD}$, UH ≥ 25 m^2/s^2	0.669	0.602	0.566	0.466	0.320
$\overline{POD}$, UH ≥ 35 m^2/s^2	0.500	0.418	0.372	0.295	0.213
$\overline{POD}$, UH ≥ 50 m^2/s^2	0.310	0.238	0.207	0.143	0.093
$\overline{POD}$, UH ≥ 75 m^2/s^2	0.132	0.088	0.072	0.043	0.016
$\overline{SR}$, UH ≥ 25 m^2/s^2	0.205	0.218	0.219	0.236	0.402
$\overline{SR}$, UH ≥ 35 m^2/s^2	0.276	0.284	0.320	0.338	0.493
$\overline{SR}$, UH ≥ 50 m^2/s^2	0.409	0.480	0.520	0.547	0.677
$\overline{SR}$, UH ≥ 75 m^2/s^2	0.647	0.642	0.682	0.764	0.868
$\overline{CSI}$, UH ≥ 25 m^2/s^2	0.186	0.191	0.187	0.186	0.217
$\overline{CSI}$, UH ≥ 35 m^2/s^2	0.216	0.203	0.208	0.187	0.174
$\overline{CSI}$, UH ≥ 50 m^2/s^2	0.214	0.189	0.174	0.127	0.089
$\overline{CSI}$, UH ≥ 75 m^2/s^2	0.123	0.084	0.070	0.042	0.016
$\overline{POD_{weighted}}$, UH ≥ 25 m^2/s^2	0.807	0.755	0.727	0.655	0.585
$\overline{POD_{weighted}}$, UH ≥ 35 m^2/s^2	0.651	0.572	0.529	0.455	0.398
$\overline{POD_{weighted}}$, UH ≥ 50 m^2/s^2	0.448	0.375	0.336	0.262	0.199
$\overline{POD_{weighted}}$, UH ≥ 75 m^2/s^2	0.233	0.172	0.141	0.087	0.046
$\overline{SR_{weighted}}$, UH ≥ 25 m^2/s^2	0.407	0.409	0.405	0.386	0.318
$\overline{SR_{weighted}}$, UH ≥ 35 m^2/s^2	0.457	0.459	0.455	0.417	0.365
$\overline{SR_{weighted}}$, UH ≥ 50 m^2/s^2	0.523	0.534	0.533	0.524	0.547
$\overline{SR_{weighted}}$, UH ≥ 75 m^2/s^2	0.631	0.615	0.669	0.691	0.791
$\overline{CSI_{weighted}}$, UH ≥ 25 m^2/s^2	0.371	0.361	0.352	0.321	0.260
$\overline{CSI_{weighted}}$, UH ≥ 35 m^2/s^2	0.367	0.342	0.323	0.278	0.235
$\overline{CSI_{weighted}}$, UH ≥ 50 m^2/s^2	0.318	0.283	0.260	0.212	0.171
$\overline{CSI_{weighted}}$, UH ≥ 75 m^2/s^2	0.205	0.156	0.132	0.084	0.046

Appendix C

Mean Contingency Table Statistics, Variations by Time of Day
(Minimum Fractional Coverage Threshold).

Quantity	12-18 UTC	18-00 UTC	00-06 UTC	06-12 UTC
$\overline{POD}$, UH ≥ 25 m^2/s^2	0.675	0.683	0.671	0.635
$\overline{SR}$, UH ≥ 25 m^2/s^2	0.214	0.175	0.233	0.209
$\overline{CSI}$, UH ≥ 25 m^2/s^2	0.194	0.162	0.209	0.187
$\overline{POD_{weighted}}$, UH ≥ 25 m^2/s^2	0.723	0.837	0.808	0.835
$\overline{SR_{weighted}}$, UH ≥ 25 m^2/s^2	0.424	0.421	0.391	0.392
$\overline{CSI_{weighted}}$, UH ≥ 25 m^2/s^2	0.365	0.389	0.357	0.364
$\overline{POD}$, UH ≥ 35 m^2/s^2	0.479	0.507	0.567	0.420
$\overline{SR}$, UH ≥ 35 m^2/s^2	0.239	0.266	0.327	0.262
$\overline{CSI}$, UH ≥ 35 m^2/s^2	0.190	0.211	0.261	0.192
$\overline{POD_{weighted}}$, UH ≥ 35 m^2/s^2	0.598	0.628	0.680	0.702
$\overline{SR_{weighted}}$, UH ≥ 35 m^2/s^2	0.460	0.460	0.461	0.441
$\overline{CSI_{weighted}}$, UH ≥ 35 m^2/s^2	0.351	0.361	0.379	0.371
$\overline{POD}$, UH ≥ 50 m^2/s^2	0.297	0.295	0.391	0.235
$\overline{SR}$, UH ≥ 50 m^2/s^2	0.431	0.340	0.463	0.428
$\overline{CSI}$, UH ≥ 50 m^2/s^2	0.213	0.187	0.269	0.179
$\overline{POD_{weighted}}$, UH ≥ 50 m^2/s^2	0.444	0.401	0.458	0.527
$\overline{SR_{weighted}}$, UH ≥ 50 m^2/s^2	0.612	0.498	0.511	0.500
$\overline{CSI_{weighted}}$, UH ≥ 50 m^2/s^2	0.346	0.286	0.318	0.345
$\overline{POD}$, UH ≥ 75 m^2/s^2	0.093	0.134	0.198	0.079
$\overline{SR}$, UH ≥ 75 m^2/s^2	0.654	0.588	0.650	0.733
$\overline{CSI}$, UH ≥ 75 m^2/s^2	0.089	0.122	0.179	0.077
$\overline{POD_{weighted}}$, UH ≥ 75 m^2/s^2	0.194	0.197	0.230	0.346
$\overline{SR_{weighted}}$, UH ≥ 75 m^2/s^2	0.715	0.630	0.605	0.590
$\overline{CSI_{weighted}}$, UH ≥ 75 m^2/s^2	0.180	0.176	0.200	0.279