UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IF WE FORECAST IT, THEY MAY (OR MAY NOT) USE IT: SUB-DAILY SEVERE WEATHER TIMING INFORMATION AND ITS UTILITY FOR FORECASTERS, STAKEHOLDERS AND END USERS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

MAKENZIE JO KROCAK
Norman, Oklahoma
2020

IF WE FORECAST IT, THEY MAY (OR MAY NOT) USE IT: SUB-DAILY SEVERE WEATHER TIMING INFORMATION AND ITS UTILITY FOR FORECASTERS, STAKEHOLDERS AND END USERS

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Harold E. Brooks, Chair

Dr. Joseph T. Ripberger

Dr. Cameron Homeyer

Dr. Elinor Martin

Dr. Jason Furtado

Dr. Justin Reedy

Acknowledgments

This dissertation is the culmination of years of work with unwavering support from countless individuals. My committee members, including Drs. Cameron Homeyer, Elinor Martin, Jason Furtado, and Justin Reedy, have shown admirable dedication and perseverance during the entire PhD process and especially during the unprecedented COVID-19 pandemic. I am truly thankful.

I have had the pleasure of working with Drs. Harold Brooks and Joe Ripberger for the entirety of my graduate career. What started as a ten-week summer internship in 2014 turned into nearly five years of graduate school and learning how to be a scientist from them. Harold and Joe, thank you for being scientists who challenge me to grow while also offering ample support when I need it. I am especially grateful for the near-daily interactions with both of them and for the flexibility and support they have shown me during my graduate school tenure. Finally, they have been extremely gracious and forgiving advisers throughout the last few months of this process, for which I am particularly thankful for.

A sincere thank you goes out to all of the teachers and professors who have inspired me to always be curious and ask questions. Kelly Brigham, Karen Engel, Sara Falkofske, and Amy Fredrickson are top-notch math and science teachers who show young women like myself that STEM careers are indeed possible for them. I am incredibly thankful for the confidence, determination, and love for science they instilled in me during my high school years. At Iowa State University, I have Dave Flory to thank for allowing me to dream big while also keeping me grounded. The close-knit department allowed me to discover meteorology without the intimidation of a large department with dozens of weather nerds. Then when I moved to a large department at the University of Oklahoma for graduate school, I was lucky enough to have numerous dedicated professors and mentors, including Alan Shapiro, Steven Cavallo, Jeff Basara, Petra Klein, and Jens Redemann. Thank you for your dedication to your students. I have always said that graduate students are treated

like colleagues at OU, and it has been my greatest pleasure to work with and learn from all of you. Finally, no one would get degrees without the staff in the School of Meteorology and the Cooperative Institute for Mesoscale Meteorological studies, including Tracy Reinke, Christie Upchurch, Tanya Riley, and Jamie Foucher. Thank you for your support and guidance during this entire process and for being invested in me as a student and as a person.

In addition to the School of Meteorology community, I have had the privilege to be a part of the crew at the OU National Institute for Risk and Resilience. Hank Jenkins-Smith, Carol Silva, Kuhika Gupta, Nina Carlson, and everyone at the institute has shown me what a work family truly is. Thank you for the inspiration, lunch time conversations, and personal investment you make in all of us.

I am also grateful for the army of friends who have made this process rewarding and immensely enjoyable, even in the most challenging times. Rachel Miller, Lauren Walker, Allie Brannan, Addison Alford, Jackson Anthony, Kat Gebauer, Josh Gebauer, Kristen Perez-Rickels, Drew Rickels, Heather Wade, Andy Wade, Jinan Allan, Sean Ernst, Wesley Wehde, Burkely Gallo, and dozens of others are the most patient and gracious friends I could have asked for. From cheese fries and swirls to brewery tours and mid-general exam dinner nights, I am so grateful for all the moments we have shared that have made graduate school my favorite years thus far. Also, thank you to Lans Rothfusz, Kathy Rothfusz, Gerry Creager, and Bari Creager for providing advice and support throughout my time as a graduate student. Being so far from family can be difficult, but you always made sure we had company for dinner and a place to spend the holidays. For that, I am thankful.

My family also deserves a hearty ”you did it!”, as I can promise I would not have even considered a graduate degree without their support. Many parents may have rolled their eyes at a strong-willed 17-year-old confidently stating that they were going to go to school for meteorology and would get it done in three years. Mom and Dad, thank you for always fostering my curiosity and raising me to believe that I can accomplish whatever

I set out to do. Sean and Maddie, thank you for being a constant source of laughter and annoyance, and for choosing to be my best friends. Sean, THIS is the graduation ceremony you should come too.

Finally, my meteorology education has led me to multiple degrees and also a life-long partner. Matt, you are a constant source of support, curious science questions, and unconditional love. Thank you for growing with me as a scientist and as a partner. Thank you for learning when to be my strongest advocate and when to stand to the side and let me advocate for myself. Thank you for being exactly whom I need, when I need you. Lastly, to Diego and Mogli, thank you for being ridiculous dog-cats, providing a consistent source of love and angst.

Table of Contents

Acknowledgments **iv**

List of Tables **ix**

List of Figures **x**

Abstract **xv**

1 Introduction **1**

2 An Analysis of Subdaily Severe Thunderstorm Probabilities for the United States **5**

2.1 Introduction and Background 5
2.2 Data and Methods 7
2.3 Results 9
2.4 Discussion 16

3 Evaluating the creation and assumption of use of forecasted timing products **20**

3.1 Introduction and background 20
3.2 The Hazardous Weather Testbed 24
3.2.1 The Spring Forecasting Experiment 24
3.2.2 The end user experiments 25
3.3 Initial development 28
3.4 The first iteration: isochrones in the 2016 and 2017 textbed experiments . . 31
3.4.1 Data collection 31
3.4.2 Lessons learned 33
3.5 The second iteration: PSTs in the 2018 testbed experiments 40
3.5.1 Data collection 40
3.5.2 Lessons learned 45
3.6 Continued refinements: the 2019 testbed experiments 54
3.6.1 Data collection 54
3.6.2 Lessons Learned 63
3.7 Discussion 82

4 The impact of hours of advance notice on protective action in response to tornadoes **86**

4.1 Introduction and background 86
4.2 Data and methods 89
4.2.1 Survey data 89
4.2.2 Response treatment 91

4.3 Results and Discussion 93
4.3.1 Word analysis 93
4.3.2 Categorical Analysis 96
4.4 Summary and conclusions 99

5 Discussion 103

Reference List 107

Appendix A

Survey instruments 112
A.1 Sample SFE forecaster survey questions 112
A.1.1 Daily forecast evaluation 112
A.1.2 End of week evaluation 114
A.2 Sample end user survey and focus group questions 116
A.2.1 Pre-test 116
A.2.2 Mid-experiment micro-survey 119
A.2.3 Post-test 119
A.2.4 EM focus group questions 125

Appendix B

Emergency Manager Participants 127

List of Tables

3.1 Example of 2018 EM experiment schedule 26

3.2 Example of 2019 EM experiment schedule 28

3.3 Lessons learned in the 2016 and 2017 SFE experiments and future actions to consider. 40

3.4 Lessons learned in the 2018 experiments and future actions to consider. . . 54

3.5 Lessons learned in the 2019 experiments and future actions to consider. . . 81

4.1 Response categories and their descriptions. 93

B.1 2018 emergency manager participants 127

B.2 2019 emergency manager participants 128

List of Figures

2.1 An example of a daily time series of reports for a single location. The green shaded area represents six hours, the yellow represents four hours, and the red represents one hour of the day. The percentages reflect the fraction of reports captured in each timeframe for this particular example. 9

2.2 The percentage of reports not captured (y-axis) at differing time periods (x-axis) within the 24-hour convective outlook day. The percentages are expressed on a logarithmic scale to show detail at the smallest values. The dashed line indicates 5% of reports not captured. 10

2.3 The percentage of all daily reports within 40-km of a point captured in a 4-hour period of a 24-hour convective outlook day (12 UTC - 1200 UTC). Data is reported for grid points with at least 20 reports over the 65-year study period. 12

2.4 The percentage of all daily reports within 40-km of a point captured in a 4-hour period for Norman, OK; Huntsville, AL; Columbus, OH; Des Moines, IA; Raleigh, NC; and Denver, CO. 13

2.5 The median daily start time (in UTC) of the 4-hour period that captures the highest percentage of the daily severe weather reports within 40-km of a point. Data is reported for grid points with at least 20 reports over the 65-year study period. The dashed line indicates the delineation between the eastern and central US used in Figure 2.6. 14

2.6 Overlapping histograms of the daily start times of the 4-hour period that captures the most daily reports within 40-km of a point for roughly 65-90 degrees longitude west and for 91-105 degrees longitude west. 16

3.1 One of the original visualizations for a timing information product featuring lines that represent the start time of the 4-hour period forecasted to contain the majority of daily local storm reports. 29

3.2 One of the original visualizations for a timing information product featuring boxes that represent the time frame of the 4-hour period forecasted to contain the majority of daily local storm reports. 30

3.3 A schematic of isochrone activities in the 2016 and 2017 SFE. 32

3.4 Isochrone verification plots showing the forecasted time frames (left), the observed time frames (middle), and the observed minus the forecasted time frames (right). 34

3.5 A histogram of the aggregated isochrone verification shown in Figure 3.4 for the 2016 SFE. 36

3.6 A histogram of the aggregated isochrone verification shown in Figure 3.4 for the 2017 SFE. 37

3.7 Examples of of the probabilistic total severe weather forecasts (left) and the practically perfect probabilistic forecasts (right) for notable cases in the 2018 SFE. Green dots are hail reports (greater than 1 inch reports are black dots), blue squares are wind reports (over 65 knots are black squares), and red triangles are tornado reports (EF2 and greater reports are black triangles). 42

3.8 An example of the simulated PST product shown to EMs during the 2018 experiment. 45

3.9 An example of forecasted PST areas drawn by participants during the 2018 SFE. 47

3.10 An example of forecasted PST areas drawn by participants during the 2018 SFE. 48

3.11 A performance diagram showing the probability of detection (y-axis) and the success ratio (x-axis) of the 2018 SFE PSTs (blue) and operational weather watches issued during the same time period (red). Each dot represents a single day, and the larger square is the mean performance. 50

3.12 Examples of of the probabilistic total severe weather forecasts (left) and the practically perfect probabilistic forecasts (right) for notable cases in the 2019 SFE. Green dots are hail reports (greater than 1 inch reports are black dots), blue squares are wind reports (over 65 knots are black squares), and red triangles are tornado reports (EF2 and greater reports are black triangles). 55

3.13 An example of of the probabilistic total severe weather forecast (left) and the PST timing areas (right) that were drawn in the 2019 SFE. 56

3.14 The probability density function of the probability of detection for automatic PST areas generated using three different thresholds (noted as a percentage of the maximum grid value). 58

3.15 The probability density function of PST area for three different thresholds (noted as a percentage of the maximum grid value). 59

3.16 An example of a simulated PST for 19 July 2018 used in the 2019 end user experiment. 62

3.17 Forecaster response to the survey prompt: ”How much do you agree or disagree with the following statement: The added value of the PST product is greater than the added workload”. 66

3.18 Forecaster response to the survey question: ”How difficult was it to create the PSTs yesterday?” 67

3.19 Forecaster response to the survey question: ”Was the ”first guess PST” guidance product from your ensemble subset useful?” 68

3.20 Forecaster response to the survey question: ”Who do you think the audience for the PST products should be?” Respondents were allowed to check more than one option. 71

3.21 End user response to the survey question: Please evaluate the degree to which the PST: - Was easy to use”. 72

3.22 End user response to the survey question: ”Please evaluate the degree to which the PST: - Made you more confident in your decisions”. 73

3.23 End user response to the survey question: ”How much do you agree or disagree with the following statement? I believe I would use the PST on most severe weather days for my area”. 74

3.24 The figure shown during post-test surveys discussing overlapping PST areas. The question wording was: ”Below is an example of a forecasted PST area. Please indicate which timeframe is forecasted for the red star.” 80

3.25 A summary of the timing product evolution throughout the testing period from Spring 2016 through winter 2019. 84

4.1 The percentage and number of responses containing the most common words in response to 4 hours of advance notice (a, n=1392) and 8 hours of advance notice (b, n=1404) for a dangerous tornado. 94

4.2 Difference in percentage of use between 8 hours (orange bars indicate higher percentage use) and 4 hours (blue bars indicate higher percentage use) of advance notice. * p $<$ 0.1. 95

4.3 The percentage and number of responses in each response category for 4 hours (a, n=1392) and 8 hours (b, n=1404) of advance notice. 97

4.4 The difference in percentage of response categories between 8 hours (orange bars indicate more responses in this category) and 4 hours (blue bars indicate more responses in this category) of advance notice. * p $<$ 0.1. . . . 98

A.1 An example PST graphic. 115

A.2 An example PST graphic. 124

Abstract

Severe weather causes significant damage and disruption to daily lives, particularly in the eastern half of the United States. While warning creation and communication has been studied extensively, much less work has been devoted to understanding weather information effectiveness on sub-daily scales that are longer than the warning scale. This work focuses on understanding if and how scientists can provide timing information on the convective outlook scale (4-24 hours from the event), and whether that information would be useful to stakeholders and the general public. A mixed-method approach was used, including interviews, surveys, and focus groups with forecasters, emergency managers, and the US population.

Results from the initial feasibility study indicate that a majority of daily severe weather reports at a single location are concentrated within short (around four hours) periods of the convective day. This result is consistent across the US, indicating that timing information could be included within the convective outlook without a drastic change in the product definitions. However, initial testing with forecasters showed that forecasting these shorter periods of time was sufficiently challenging and would need in-depth training and improved verification and calibration methods. Although these challenges did exist, forecasters still noted the value they see in this type of product. Emergency managers also found timing information to be extremely valuable to their decision-making process. They note that the timing of severe weather events impacts when, how, and where they stage resources and dictates what different entities (like schools, workplaces, and municipal services) do to prepare. Finally, respondents to a national survey of US residents generally reported that they would monitor the weather and make preparations if they were given knowledge of specific timing of severe weather for their area. Many of these respondents also said they would shelter hours before the storms were forecasted to arrive, indicating that future work needs to be done to minimize anchoring effects that are likely occurring during these hypothetical severe weather scenarios.

Ultimately, this work is an attempt to involve researchers, practitioners, and end users in the process of developing new forecast information. The co-production of scientific knowledge and a formal development, testing, refinement, and implementation process will become more important as NOAA implements a more continuous flow of weather information. Such changes must ultimately be vetted with the people and organizations that rely on the weather information we provide. To produce the most useful and effective information, scientists, funding agencies, and policy makers must be flexible enough to allow researchers to see the testing and refinements of scientific knowledge as an asset to the process, not an obligation.

Chapter 1

Introduction

Severe weather events in the United States (US) are one of the costliest and deadliest natural events every year. The mean cost of severe weather events between 1980 and 2019 was 6.2 billion dollars per year, with an average death rate of 41 deaths per year (NOAA National Centers for Environmental Information 2020). Beyond the risk to the general US population, there are certain populations that are more vulnerable to severe weather events than others. For example, residents in the southeastern US (SE US) experience a higher death rate due to tornadoes than the traditional tornado alley, likely because of the unique overlap in tornado occurrence, population density, and higher proportions of residents living in poverty and mobile homes (Ashley 2007). Beyond the economic vulnerabilities examined in Ashley (2007), other populations, like those who are disabled, live in rural areas, the elderly, or those without access to an adequate storm shelter are also particularly vulnerable to severe weather events.

The current severe weather notification system utilized by NOAA and the National Weather Service (NWS) consists of discrete product levels. On the event scale, the products start generally 8 days out with the Storm Prediction Center (SPC) convective outlook product. This product is issued 4 times per day, is valid for 24 hours, and includes information about the forecasted amount of severe weather and what types (hail, wind, or tornado) are most likely. Then on the day of the severe weather event, the SPC can issue weather watches, which are usually valid from a few hours before the event until the event is over. The SPC forecasts a severe weather event within 25 nautical miles of a point, so any report within 25 nm of a point with forecasted probabilities will verify the SPC forecast. At a more local level, NWS offices issue severe thunderstorm and tornado warnings, which are usually valid from 30 minutes (or generally much less) before the event until the event is over. As it currently stands, there is no formal forecast information about the timing of a severe weather threat issued until the weather watch comes out (presumably a few hours before the event, when people have already gone to work or started their day). This can be particularly problematic for the vulnerable populations described above. Without a plan in place to stage people in locations with easily accessible shelters, the few minutes to a few hours of notification is likely inadequate to properly shelter many individuals. Decision makers also need more time to make plans, or they need information at critical decision points, which the current severe weather notification system does not formally provide.

One possible solution to this lack of information at critical decision points is to develop a system that provides a more continuous flow of weather information. In theory, this would allow a wide variety of decision makers (with unique decision points) to have access to information whenever they needed it. This research initiative, currently under way within NOAA, is known as the forecasting a continuum of environmental threats, or FACETs project. The goal of this work is to create a system that is less dependent on the current discrete product levels and more continuous in providing weather information (Rothfusz et al. 2018). While much of the FACETs work to date has been focused on implementing probabilistic hazard information on the warning scale (e.g. Ling et al. 2015), the vision for this project is to extend the continuous flow of information out to days before the event (i.e. the SPC scale), which would also benefit our most vulnerable populations who need more time than what the warning scale provides to prepare for events. However, it is also important to note that the scheduled structure of the current system means that decision makers can expect information at certain times of the day. This is an artifact that may have to remain in place, or potentially built upon, in a new system.

There is one particularly important piece of information that is not explicitly provided in the SPC convective outlook product; the forecasted timing of severe weather. Timing information has been shown to be important during the decision making process. Reed and Senkbeil (2019) found that the timing and intensity of weather events were the most important pieces of information for members of the public. They found this by surveying the public about the extended forecast graphics used by broadcast meteorologists. Unfortunately, forecasting the timing of severe weather events (outside of generic terms like the afternoon and evening) is challenging on the day of the event and nearly impossible at longer periods.

There also has not been a comprehensive analysis of what response actions may look like if decision makers and the public were given timing information at longer lead times. However, there has been some work related to warning-scale response actions, including many studies that attempt to measure what fraction of participants respond to tornado warnings. Some of these studies include interviews after actual tornado events and find that anywhere from 43% to 79% of residents take action, depending on the region where the event occurred (Balluz et al. 2000, Miran et al. 2018b, Chaney et al. 2013). Other work that measures hypothetical situations find higher response rates, with anywhere from 75% to 90% of respondents claiming they will take action during a future event (Schultz et al. 2010, Lindell et al. 2016, Ripberger et al. 2015). For a more comprehensive overview of public response to tornado events, see section 4.1.

As new technology allows forecasters and policy makers to include more detailed information at relevant points in the event timeline (like, say, timing information in the convective outlook products), there needs to be a formal testing, refinement, and implementation process in place to ensure all new technology is thoroughly vetted before it becomes operational. Demuth et al. (2020) is a recent example of work that is iterative and interdisciplinary in the severe weather domain. First, researchers observed and documented forecaster needs, then developed products to try to solve those needs and tested them with forecasters. Then they went back to the developers with the feedback from forecasters to improve the products. This iterative process it vitally important to creating useful forecast information, whether that be for forecasters or for users. Ideally, this process would include both creators and practitioners. Although the co-production of scientific knowledge with stakeholder input is not a new concept, it has been relatively untested in the severe weather domain. Some work, like Pielke (1997), raise the need for physical scientists to engage with social scientists from the onset of a research endeavor. While social science has been more fully integrated into weather research over the last two decades, there is still more work to be done. A recent example of stakeholder involvement in the development of weather products comes from the tropical cyclone domain. Morrow et al. (2015) focused on improving storm surge communication by soliciting feedback from multiple user groups. They ultimately found that the feedback from broadcast meteorologists and emergency managers was particularly important because they raise important communication issues during informal discussions that the more formal experimentation would not have captured. For a more complete overview of the co-production of science within the weather domain, see section 3.1.

Ultimately, developing new technologies without the input of the intended users is unsatisfying at best and can lead to missed opportunities and years of wasted work at worst. A comprehensive research, development, testing, and implementation process should be developed to include all parties, from researchers to forecasters, stakeholders, and end users. This dissertation is an illustration of what this process could look like for the development of sub-daily severe weather information. Ultimately, we show the entire process, from analyzing if forecasting timing information is physically possible (see section 2), to understanding if (and to whom) that information is truly useful (see sections 3 and 4). The following chapters explain how this process was undertaken, while highlighting our attempt to make the entire process more interdisciplinary in nature.

Chapter 2

An Analysis of Subdaily Severe Thunderstorm Probabilities for the United States

2.1 Introduction and Background

Experts in the field of weather risk communication show that end users generally understand the existence of underlying uncertainty in weather forecast information (e.g., Morss et al. 2019, Joslyn and Savelli 2010, Savelli and Joslyn 2012, Fundel et al. 2019). Murphy (1993) also notes that for simple decision problems, non-probabilistic forecasts maximize the number of users who lose value. As such, research organizations have called for using probabilities to describe forecast uncertainty, as it may be beneficial to residents when making response action decisions (National Research Council 2006, American Meteorological Society Council 2008). Following these recommendations, NOAA is currently developing a paradigm that includes rapidly updating probabilistic information for user-specific locations. The FACETs project aims to provide a continuous stream of probabilistic information to keep people up to date on weather, water, and climate threats from days or more out down to minutes before the event occurs (Rothfusz et al. 2018).

The current NWS product structure for severe weather consists of three product levels; the convective outlook (which is issued by the SPC from one to eight days in advance for the Continental United States (CONUS)), severe thunderstorm and tornado watches (which are also issued by the SPC generally 1–4 hours before the event occurs, with a mean size of 30,000 square miles, which is about the size of Maine), and then the warning (issued by a local NWS office 0–60 minutes before the event and has a mean size of 250 square miles). One of the early challenges of the FACETs project was the reliance of community infrastructure on these current products. For example, some communities often use a specific product (like a tornado watch) to activate procedures (Cross et al. 2019). Should the new system consist of just probabilities, the previous decision points that relied on specific products would need to be changed. As work has continued in this domain, many researchers and forecasters have come to realize that some products would need to remain in place to ensure that the transition between systems was smooth.

Since transitioning from the current system will likely be evolutionary and take time and discourse, researchers have begun working on ways to provide more information within and in-between the current product levels. As part of this effort, research scientists and forecasters need to understand how the probabilities of these events change between different reference classes. A reference class refers to the event of interest. For example, the SPC forecasts the probability of an event within a 25 nautical mile radius circle. However, most people are probably more interested in knowing the likelihood of an event *at their house*. However, the probability of a tornado hitting a 25 nautical mile circle is much larger than the probability of it hitting a single point (like a house). In fact, the probability of it hitting a single point is so small that the numbers are likely meaningless to people (i.e. a 0.0001% chance of a tornado is probably not very useful information). This trade off between specific location information and meaningful probabilities presents a challenge for researchers and forecasters. This work focuses on understanding the probabilities of severe weather on a sub-daily scale, but larger than the warning scale (i.e., spatiotemporal scales between the convective outlook and watch, generally on a state to regional spatial scale and temporal scales between one and 24 hours), such that forecasters can assign correct values that are also meaningful to users.

To begin this process, we start by analyzing the distribution of events within a day at any single location. We use the general SPC convective outlook probabilities as a simple starting point. While the probabilities are forecasted for up to a 24-hour period, intuitively, many meteorologists know that at any location the probability of severe weather is actually near zero for a large portion of the day, then it increases to the forecasted probability shortly before the event begins, and then decreases back to near zero shortly after the event ends. Following this example, we define an event as a local storm report within 40 km (25 mi) of a point to match the spatial scales of the current SPC convective outlook probabilities (NOAA Storm Prediction Center 2019a). We then investigate how the events on a single day are distributed in time. Are they spread out across the day or concentrated within a smaller period of time? If there is a smaller window of time when most of the events are concentrated, when does that window start? Is there regional variability in the duration of severe reports or the start time of the smaller window of threat? Given that our analysis has nearly identical spatial scales to the current SPC convective outlook probabilities, knowledge of the climatological duration of severe weather events means the SPC convective outlook probabilities could be valid for a smaller window of the day. The forecasting challenge would then be to identify when that window starts and ends. From a communication standpoint, knowing the forecasted window of threat on a severe weather day could help the entire range of decision makers, from emergency management and school officials to youth coaches and individuals, decide how to prepare in advance of the start of the event.

2.2 Data and Methods

Hail, wind, and tornado reports from the SPC Severe Report Database (NOAA Storm Prediction Center 2019b) between 1950 and 2015 are used to calculate the spatiotemporal scales that severe events generally fit within, i.e., a spatial area and temporal duration that captures a majority of the daily events. While there are known issues with the report database, especially with regards to the increase in the number of reports (see Doswell and Burgess 1988, Trapp et al. 2006, Verbout et al. 2006), it is the most comprehensive severe weather occurrence database for the United States and we believe the data still provides useful insight into the general pattern of severe weather events.

We begin by identifying all of the reports within a specified radius (we test 10, 20, 40, 80, and 200 km) around a point in the Contiguous United States (CONUS). While we test multiple spatial scales, we focus most of our analysis on the 40 km radius so that our results could speak to the definitions of the current SPC products and allow for the testing and verification of new products within the SPC forecast domain.

Next, at each point with at least 20 reports over the 65 year period, we create time series of the reports for each convective (1200 UTC to 1200 UTC) day. An example of this time series is shown in Figure 1. Using these time series, we calculate a variety of quantities including the maximum percentage of the daily reports that are captured within smaller timeframes (specifically within 1, 2, 4, 6, 8, and 12 hours of the day), and the start time of the maximum daily window. The percentage of reports captured in smaller timeframes is calculated at each grid point as:

$$p_{captured} = \frac{\sum r_{captured}}{\sum r_{total}} \cdot 100 \tag{2.1}$$

where the numerator is the sum of the reports captured within the specified smaller timeframe on all days, and the denominator is the total number of reports that occurred at that grid point. In Figure 2.1, the numerator would be the number of reports captured in the shaded areas (green showing six hours of the day, yellow showing the four hours of the day, and red showing one hour of the day with the maximum number of reports captured), and the denominator would be the total number of reports shown in the time series. Obviously, if there is only one report at a point over the convective outlook day, then the total percentage of reports captured that day is 100%. Then the total number of reports in the numerator and denominator are aggregated over all days and all grid points. To analyze regional differences in the window start time, the timestamp of the start of the smaller timeframe is calculated for each day. Then the median start time at each grid point is calculated.

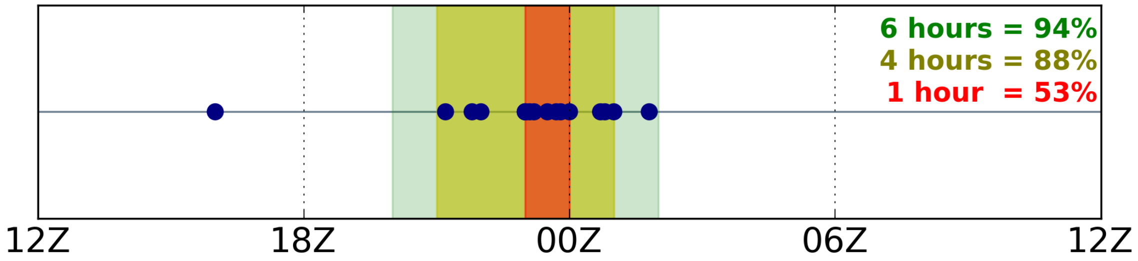


Figure 2.1: An example of a daily time series of reports for a single location. The green shaded area represents six hours, the yellow represents four hours, and the red represents one hour of the day. The percentages reflect the fraction of reports captured in each timeframe for this particular example.

2.3 Results

To begin the analysis, we aggregate all of the points across the CONUS to obtain a holistic view of the spatiotemporal scales of severe weather events. First, we analyze all of the reports within 40 km of a point for all points across the CONUS (Fig. 2.2, green line). Within any single convective outlook day, more than 99% of reports will be contained within just 12 hours of the full day. Furthermore, over 95% of daily reports within 40 km of any point occur in a 4-hour period, and a single hour of the day still captures more than 80% of the daily reports. If we consider the probability behavior of uniformly distributed points (i.e., events occurring equally across the 24-hour period), the percentage of reports captured drops to 50% at 12 hours and just 16.7% at 4 hours (Fig. 2.2, grey line). Clearly, severe weather events at any given point are concentrated in timeframes smaller than 24 hours, with a vast majority of reports occurring in just 4 hours of the day. Moreover, since the spatial scales of this analysis are nearly identical to the SPC's definition of an event (i.e., a report occurring within 25 nautical miles of a point), it follows that the SPC's probabilities at any given point can be interpreted as valid for 4 hours of the day within a reasonable margin of error (over 95% of reports versus 100% of reports).

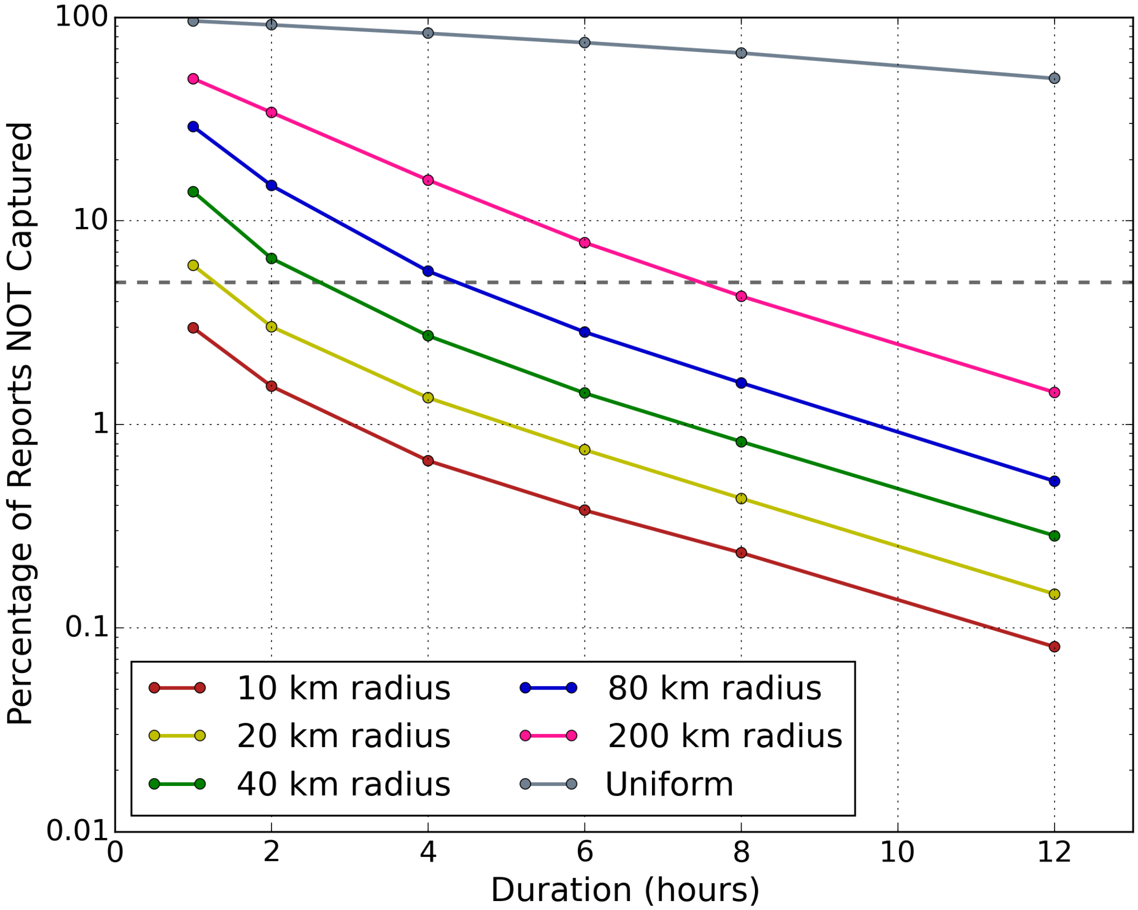


Figure 2.2: The percentage of reports not captured (y-axis) at differing time periods (x-axis) within the 24-hour convective outlook day. The percentages are expressed on a logarithmic scale to show detail at the smallest values. The dashed line indicates 5% of reports not captured.

After analyzing severe weather probabilities on varying temporal scales, we also calculate the percentage of events captured within numerous radii around a point and numerous temporal durations (Fig. 2.2). For all radii, the percentage of reports that aren't captured increases with increasing radii around a point and decreasing temporal durations. This probability behavior is largely intuitive because it takes longer for weather systems to cover a 200 km radius (similar to the north-south extent of Oklahoma) than a 40 km radius (similar to the size of the Oklahoma City limits). In other words, reports will be occurring for a longer period of time when looking at an area the size of Oklahoma versus an area the size of Oklahoma City. More than 95% of reports within 10 km of a radius are captured within a single hour of the day or longer (represented by the points below the dashed line in Fig. 2.2). Longer temporal durations are needed to capture more than 95% of daily reports at other radii. The 40 km radius needs at least four hours, and the 200 km radius needs 8 hours (Fig. 2.2).

While the main goal of this work is to understand how severe event probabilities behave within differing spatiotemporal scales, it is also critical to understand how these behaviors differ by location. To align with the current SPC definition of an event (defined as a severe weather report within 25 miles-or roughly 40 km-of a point, (NOAA Storm Prediction Center 2019a)) and still capture a majority of daily reports (see the green line in Fig. 2.2), the 40 km spatial scale and 4-hour temporal scale are used for further investigation. To this end, the percentage of reports captured is calculated for each individual point across the CONUS on an 80 km grid (Fig. 2.3). More than 90% of all reports within 40 km of a point are captured in 4 hours of the 24-hour convective outlook day for all points east of the Rocky Mountains (where most severe events occur). Therefore, the current 24-hour convective outlook probabilities forecasted by the SPC could be interpreted as 4-hour probabilities with different start times depending on the location and day. This finding is important because any products that use this definition (like a convective outlook-type product) need to have consistent probabilistic definitions of events across the entire domain. Since there are no strong gradients in probability, any future products that use this definition will remain consistent no matter where the product is placed in the CONUS. An example of an experimental product might be a convective outlook that includes the probabilities of an event occurring along with a forecasted 4-hour timeframe of when that event may occur. Since the percentages of reports that climatologically occur within 4-hours at any given point are largely the same across the entire CONUS domain (Fig. 2.3), decision makers can be sure that the product is valid no matter where severe weather is forecasted.

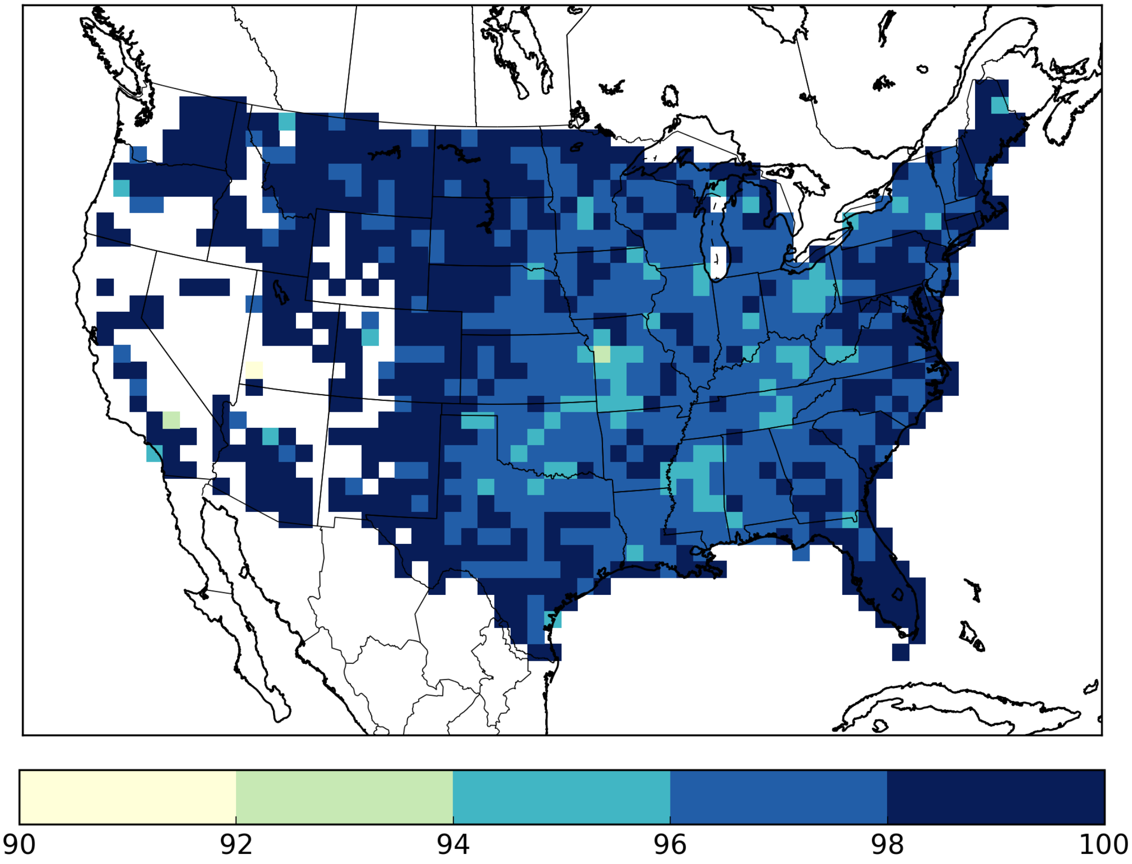


Figure 2.3: The percentage of all daily reports within 40-km of a point captured in a 4-hour period of a 24-hour convective outlook day (12 UTC - 1200 UTC). Data is reported for grid points with at least 20 reports over the 65-year study period.

In addition to understanding how severe weather events vary by region, we also investigate how event durations at a single point vary by season. The same ratio of events in a 4-hour period is calculated at each point for all 12 months. We focus on six locations (Norman, OK; Huntsville, AL; Columbus, OH; Des Moines, IA; Raleigh, NC; and Denver, CO) because they illustrate the differences between regions of the CONUS (Fig. 2.4). There is a drop in percentage of reports captured at all locations during the peak tornado season (Krocak and Brooks 2018), and then a subsequent increase afterwards. Norman and Huntsville have relative minimums during April (the spring tornado season) and September/October (the secondary fall tornado season). Similarly, Raleigh has a relative minimum in May and a second decrease in September. Columbus and Des Moines both see the minimum percentage captured in July (again, aligned with the peak in tornado occurrence for those locations). Finally, Denver has a small decrease in June, which may be in part due to sample size as well as tornado seasonality. The dips in percentages may also be explained by overnight convection trailing into the morning hours, followed by a more substantial event starting in the afternoon and evening of the following day. Some of these trends may be muted because we chose to look at the totality of severe weather reports, instead of focusing on individual hazards. While there have been some studies that examine the

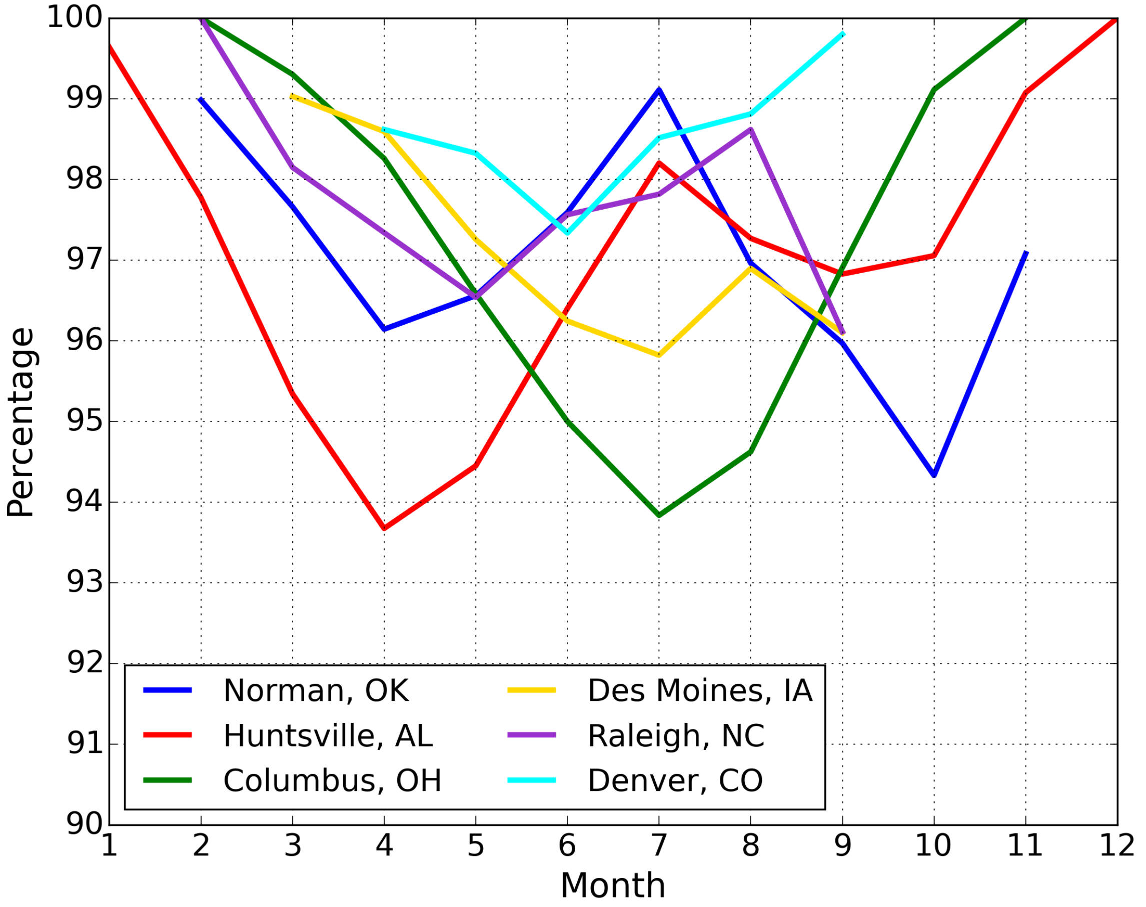


Figure 2.4: The percentage of all daily reports within 40-km of a point captured in a 4-hour period for Norman, OK; Huntsville, AL; Columbus, OH; Des Moines, IA; Raleigh, NC; and Denver, CO.

spatiotemporal patterns of tornado reports in more depth (e.g., Krocak and Brooks 2018, Brooks et al. 2003), more work needs to be done to investigate how these trends hold up when looking at hail or wind occurrence.

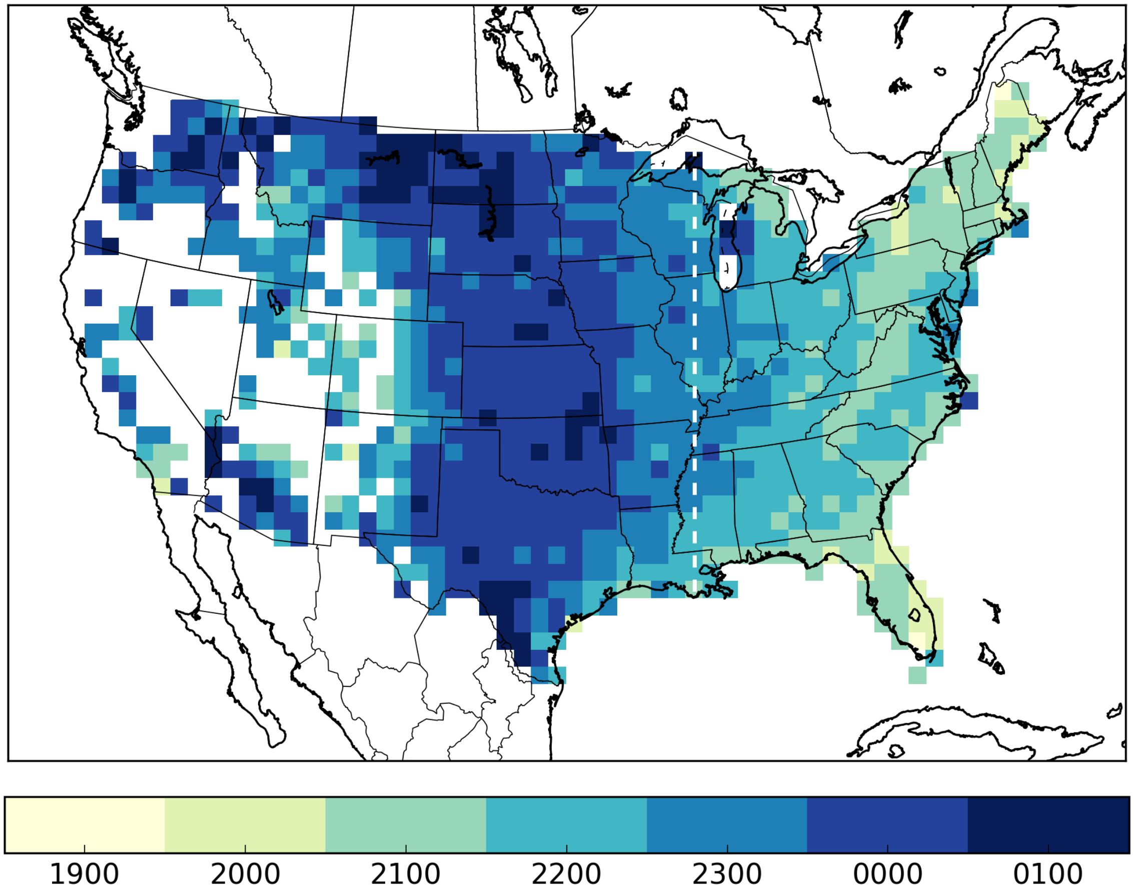


Figure 2.5: The median daily start time (in UTC) of the 4-hour period that captures the highest percentage of the daily severe weather reports within 40-km of a point. Data is reported for grid points with at least 20 reports over the 65-year study period. The dashed line indicates the delineation between the eastern and central US used in Figure 2.6.

Next, assuming that convective outlook probabilities can be interpreted as applying to smaller time periods within the day, we want to know when the climatological start time of those smaller time periods are. To accomplish this, the start time of the 4-hour period on each day with severe reports is found and then the median of all the start times is calculated at each grid point. Start times in the Central Plains are generally around 0000 UTC and become progressively earlier towards the East Coast, where start times are around 2100-2200 UTC (Fig. 2.5). In addition to the local solar time (i.e., diurnal heating) being later relative to UTC moving from east to west across the CONUS, some physical mechanisms such as the Elevated Mixed Layer (EML) inversion (Lanicci and Warner 1991), orographic lift, sea breezes, and the low-level jet may result in storms initiating later in UTC time for the Plains relative to the East Coast.

The differences in severe weather timing can be seen even more clearly when the start times of the 4-hour periods are grouped together by region (Fig. 2.6). We define the central US as the region between 91-105 degrees longitude west and the eastern region between 65-90 degrees longitude west. The entire distribution of start times is shifted later in the day when comparing the central region to the eastern region. While some of this change is due to the difference in local time and the diurnal cycle, there are still two severe weather time periods, one for the eastern part of the country starting between 2000 and 2300 UTC, and one for the central portion of the country starting between 2200 and 0200 UTC. This equates to a majority of severe weather in the eastern part of the country occurring between 2000 and 0300 UTC for any given day, and a majority of severe weather in the Plains occurring between 2200 and 0500 UTC. Regardless of location in the CONUS, these peak periods for severe weather are a good guide for potential impacts on late afternoon and evening activities and public safety.

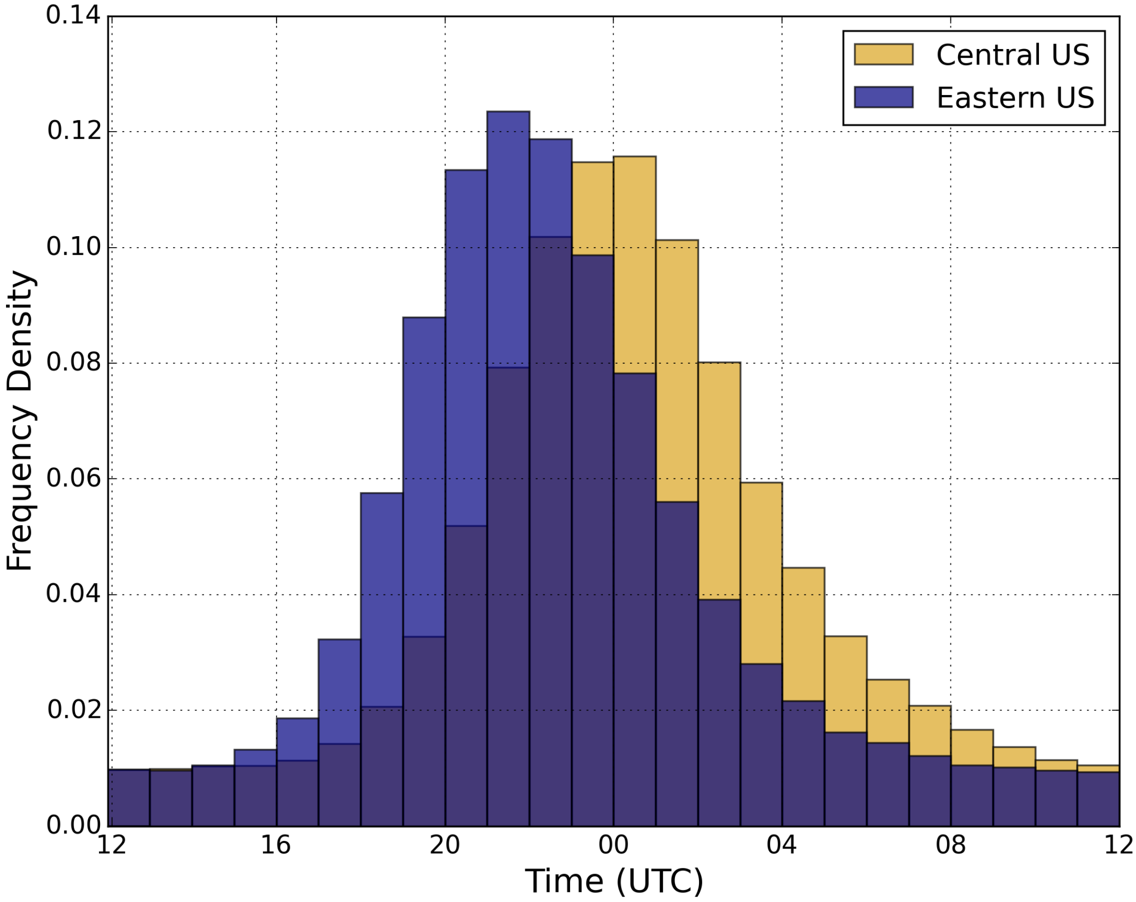


Figure 2.6: Overlapping histograms of the daily start times of the 4-hour period that captures the most daily reports within 40-km of a point for roughly 65-90 degrees longitude west and for 91-105 degrees longitude west.

2.4 Discussion

As a new generation of probabilistic severe weather products begins to take shape, researchers and forecasters are continually analyzing the best strategies for providing probabilistic information that is both accurate and useful to decision makers. This study illuminates one possibility for using probabilistic information to transition from the current hazardous weather alert system to one with higher spatiotemporal granularity and objective consistency, at least on larger spatiotemporal scales. Some of this information has already been tested with forecasters and users (e.g., Skinner et al. 2018, Wilson et al. 2019), and others are still well in the development stage.

We hope that this work serves as a foundation for future product development by analyzing the probabilities of severe weather events on spatiotemporal scales smaller than the SPC convective outlooks, but larger than warnings issued by the National Weather Service. Results show that a vast majority of daily severe weather reports at any given point occur within smaller timeframes. In fact, more than 95% of reports within 40 kilometers of a point occur in a 4-hour period, meaning that the 24-hour convective outlook probabilities assigned by forecasters in the SPC could be interpreted as 4-hour probabilities within a reasonable margin of error. If such an interpretation is considered, then the forecasting question becomes "which 4-hour period is it?" While there are many NWS weather forecast offices that offer timing information for severe weather, this is not a standardized practice and it is not required of any forecast office. If there was a standardized, regularly issued product that showed timing information for severe weather well ahead of the event start time, decision makers may be able to make informed preparedness decisions (like opening emergency operations centers, adjusting staffing levels, releasing employees or students early, etc.) with more advanced notice.

Probabilities of severe events are also analyzed spatially based on location in the U.S. Four-hour percentages of reports captured show that those ratios are consistent across all portions of the country east of the Rocky Mountains, ranging between 90 and 100%. This result is promising as products placed across different regions would have consistent definitions and probabilities of events. These percentages are also relatively consistent across seasons, with most locations seeing at least 94% of reports captured during any given month. In addition to the percentage of reports captured, the start time of the maximum 4-hour period is also analyzed spatially. The most notable trend is seen by the later start times in the Plains and earlier start times on the East Coast. The peak start time in the

Plains is around three hours later than the start times on the East Coast, although some of those differences are due to the differences in local solar time.

Ultimately, the goal of any forecasting system should be to provide users with accurate and useful information that can aid in the decision-making process. While some of the current system's product structure likely needs to remain as it is, additional information about the likelihood and timing of hazardous weather could be embedded within and in-between the current product levels. This work is meant to provide baseline knowledge of the concentration and spatiotemporal structure of severe weather events in the United States. Given that most events are concentrated within 4 hours, it is reasonable to think that forecasters could highlight this smaller timeframe so stakeholders and residents can prepare ahead of time. However, future work is needed to understand how events behave on warning scales such that forecasters can provide probabilistic information that is accurate, timely, and most importantly, useful to decision makers within the severe weather communication system.

Chapter 2 items of note:

- Severe weather events are concentrated in sub-daily time periods.
- 95% of daily reports generally occur within 4 hours, and this is true for all points in the US.
- The SPC probabilities remain mostly unchanged when forecasting for time periods between 4 and 24 hours of the day.

Chapter 3

Evaluating the creation and assumption of use of forecasted timing products

3.1 Introduction and background

In the days leading up to (and during) a severe weather event, there are numerous individuals that participate in the flow of weather information. First, forecasters create information (a forecast) and issue it at a scheduled time. Then, intermediaries (like emergency managers) use that information to make decisions and communicate it to other officials and the general public. Finally, end users receive that information and then make decisions about how to proceed.

After exploring the meteorological mechanisms driving forecasters' ability (or inability) to provide timing information for convective events, we explore *how* this information could be provided *by* forecasters *to* stakeholders (namely emergency managers) so they could potentially make preparedness decisions earlier in the event timeline. In other words, these experiments evaluated timing information with the first two steps of the communication chain mentioned in the previous paragraph. To evaluate the usefulness of timing information, feedback is solicited during experiments in the NOAA Hazardous Weather Testbed (HWT), which include either live weather exercises or simulated events. We ask forecasters to create timing products and emergency managers to describe their workflow and decision making process given prototype timing information. Participant feedback from both groups is critical to the development of the timing product we describe in the following sections. The process described is purposefully iterative. We, as researchers, develop what we think would be the most efficient product to create and verify, which (perhaps unsurprisingly) is not ideal for communicating information quickly and easily. We spent four years in the experimental process with forecasters and two years (including three experiments) with emergency managers. In the following sections, we attempt to describe the creation and testing process, why both the successes and mistakes are valuable, and how the product would be different had we not embarked on this co-production process with researchers, forecasters, and stakeholders.

While the co-production of knowledge with stakeholder input is not a new concept, it has been relatively untested in the severe weather domain. Pielke (1997) discuss the need for physical scientists to engage with social sciences from the onset of a research endeavor, and indeed social science has been more fully integrated into weather research over the last two decades. Given how important meteorological knowledge is to society, the National Oceanic and Atmospheric Administration and the National Academies of Sciences also emphasize the need for collaborative work between the physical and social sciences in the weather enterprise (NOAA 2015, National Academies of Sciences et al. 2018). However, truly co-produced weather knowledge or products is still relatively rare, potentially because of how in-depth the process is. The fundamental goals of co-production processes are to be interdisciplinary, include stakeholder participation, and ultimately produce knowledge that can be demonstrated as being useful (and used) by stakeholders (Lemos and Morehouse 2005). A recent example of stakeholder involvement in the development of weather products comes from the tropical cyclone domain. Morrow et al. (2015) focused on improving storm surge communication by soliciting feedback from forecasters, broadcast meteorologists, emergency managers, and the public. They ultimately found that the feedback from broadcast meteorologists and emergency managers was particularly important because they raise important communication issues during informal discussions that the more formal experimentation would not have captured. Other studies also note that stakeholders need to be involved in the entire research process, from the framing of the problem or research question, to analyzing the results, taking actions, and even in funding decisions (e.g. Mauser et al. 2013, Kloprogge and Van Der Sluijs 2006, Greenwood and Levin 2006, Cash et al.

2006, Cornell et al. 2013). Since this is likely a logistics nightmare for many academic and government institutions, Greenwood and Levin (2006) describe a more hybrid approach, where researchers serve as facilitators, but community members are still in control of the resulting actions or decisions. Meadow et al. (2015) also describes different levels of co-production models, from contractual, where there is a unidirectional flow of information from researchers to stakeholders, to collegial, where research is undertaken at the local level and stakeholders are involved in all aspects of the project. This work is situated somewhere between the Greenwood and Levin (2006) approach and the collaborative model described in Meadow et al. (2015). Ultimately, the goals of this work are to include stakeholder input from the onset of testing and design to produce a product that contains useful information, and more importantly, would actually be used by stakeholders.

Forecasters should be included in the design and testing process because they are clearly important to the creation of forecast information on a daily basis. They follow strict issuance schedules to ensure information is getting to the right people in a timely manner. While we know that it is climatologically possible to identify a 4-hour timeframe for every location when most severe weather will occur on a day, this task may be sufficiently challenging and time-consuming that it is too costly to provide. On the other hand, this information may be deemed critical enough that forecasters are willing to adjust other time commitments to create a timing product. These are just a few of the questions and trade-offs that we explore with forecasters. A recent example of forecaster involvement in the development of new products comes from Demuth et al. (2020). In this work, researchers took an iterative approach, where they first observed the routines of local NWS forecasters, interviewed them to identify product gaps, then developed products to help fill those gaps, tested the products with forecasters, and then went back to developers to tweak the products, and so on. The iterative nature was particularly important to ensure the product was actually useful to forecasters' routines.

Emergency managers (EMs) are vitally important to the preparation and response or recovery efforts during hazardous events, including severe weather. They often coordinate with local officials (like police and fire departments), storm spotters, and city managers within the laws and operating procedures of their jurisdictions. The educational and experiential backgrounds of EMs, the information they use, and the actions and decisions they make are incredibly diverse and often dependent upon things like jurisdiction, population size, and community assets. Due to this diversity, it is important to understand that there may never be a perfect product for all EMs (let alone all stakeholders). However, the wider the population that we can share ideas with, the more diverse feedback we can incorporate into new iterations of prototype products.

Regardless of the differences in EM populations, there has been some work that looks at how EMs use weather information. EMs make many preparation decisions ahead of the warning timescale. Demuth et al. (2012) found that coastal EMs used the timing of tropical storm force winds to make decisions ahead of any watch or warning issuance. In the severe weather domain, Baumgart et al. (2008) looked at warning information and found that EMs sought tornado warning information and verification of impacts to make response decisions. On longer time scales, we want to ascertain *if* EMs would use severe weather timing information and *how* they could incorporate this information into their routine.

Since previous work focused on understanding the climatological structure of severe weather reports on an 80km grid (the same as SPC forecast products, Krocak and Brooks 2020), the prototype products that we test are meant to accompany an SPC outlook. Currently, SPC convective outlooks contain information about where the threat is, how many reports we can expect to receive, and what kind of reports we should expect (between hail, wind, and tornado reports, NOAA Storm Prediction Center 2019a). What is not formally part of the forecast is any type of timing information outside of the 24-hour period that the forecast is valid for. In theory, all of the products described in this chapter would complement the SPC convective outlook by providing timing information to a given risk area.

3.2 The Hazardous Weather Testbed

The (HWT) is located in the National Weather Center (NWC) on the University of Oklahoma's research campus in Norman, Oklahoma. The facilities in the NWC are unique in that there are multiple university, state government, and federal organizations housed in one building. NOAA has multiple organizations in the NWC, including the Norman National Weather Service weather forecast office, the Storm Prediction Center, and the National Severe Storms Laboratory (NSSL). This conglomerate of offices allows researchers and operational forecasters to work together to develop and test new forecasting technologies and techniques. These opportunities have been formalized as the HWT, where researchers develop experiments to test new forecasting technologies with developers, operational forecasters, and stakeholders. While there are numerous experiments that occur year-round in the testbed, this work will focus on two in particular; the Spring Forecasting Experiment (SFE) and the End User experiments.

3.2.1 The Spring Forecasting Experiment

The SFE occurs annually in the late spring (late April to early June). This particular experiment is a collaboration between the SPC and the NSSL that focuses on multi-day and day-of forecasts down to (but not including) the warning scale. Experiments generally include the testing of new ensemble model configurations and the utility of convection-allowing ensemble models (Clark et al. 2016, Clark et al. 2017, Gallo et al. 2018, Clark et al. 2019). Along with the testing of new guidance, this experiment also focuses on new forecasting techniques, like individual hazard forecasts or conditional intensity forecasts.

Participants in the SFE partake in the experiment for a full work week and are contributing to activities throughout the morning and afternoon. Traditionally, a typical day in the experiment starts with evaluating the forecasts made the previous day (e.g. Clark et al. 2019). After finishing the subjective evaluations, the participants break into small groups to complete a hand-analysis of upper air and surface observation maps. After the map discussion, participants break into two groups; one that focuses on forecasting individual hazards, and the other focuses on forecasting any severe weather. These different foci lend themselves to the testing of different products depending on the nature of the technique. On the total severe desk (where the timing product was tested), participants focus on producing a day one forecast before lunch and a day two forecast in the afternoon, along with other experimental techniques and model evaluations.

During experiment activities, participants either work in groups to create a consensus forecast, or they work individually on provided laptops. If they were on the laptops, they used an online forecasting tool that was developed at the Cooperative Institute for Mesoscale Meteorological Studies for use in the SFE. The tool collates output from numerous ensemble models into a single platform, which allows the experiment participants to easily compare different models, model runs, and guidance products. This comparison allows researchers to ask more specific questions than previous years about each forecaster's individual process. This individual feedback is critical to assess the strengths and weaknesses of prototype products within different forecast processes.

The total severe desk in the SFE has a lead forecaster who is in charge of leading discussions, providing guidance products and interpretation, and creating a "first attempt" forecast. Since this individual is so involved in the forecasting process, their personal routine and model preferences permeate the final forecast output. The lead forecaster position was filled by three different individuals between 2016 and 2019, meaning that there was a wide array of forecast process preferences displayed between the experiments.

3.2.2 The end user experiments

The end user experiments are a unique example of interdisciplinary work within the weather domain that provide a rare opportunity to gather feedback in an authentic and effective manner. These experiments attempt to create a simulated environment for emergency managers to see experimental products functioning in quasi-real time and provide input on everything from the initial concept to the visualization and potential use in their daily workflow. These experiments are also unique in that they are very fluid. The experimental products tested, the way in which they are tested, and the data gathered from the experiments changes dramatically depending on what organization is funding the research, what project has actually received the funding, and where the experimental products are in the development process. The timing product was tested in three separate end user experiments: spring 2018, spring 2019, and fall 2019. Each experiment had slightly different goals, and the data collected from each experiment is a reflection of those goals as well as the continued refinement of the timing product.

Table 3.1: Example of 2018 EM experiment schedule

Time	**Activity**
8:30a	Weather briefing
8:40a	Discussion and survey on Convective outlooks, watches, and PST
9:00a	Displaced simulated event 1
10:30a	Break
10:45a	Survey and discussion
11:00a	Group debrief
11:30a	Live weather briefing and discussion
12:30p	Lunch
1:30p	Weather briefing
1:40p	Discussion and survey on Convective outlooks, watches, and PST
2:00p	Displaced simulated event 2
3:30p	Break
3:45	Survey and discussion
4:00p	Group debrief
4:30p	Adjourn

Although there are many aspects of the experiment that were fluid, some basic infrastructure and procedures remained constant. All of the end user experiments were held in the NWC under the HWT umbrella. Participants are provided with the software to be tested, supplies to record notes or thoughts about the products, and an environment conducive for group discussion and questions. The experiments lasted either two or three weeks in total with participant numbers ranging from 8 to 11. Each experiment also included a warning timescale exercise where researchers tested a new warning paradigm that includes probabilistic warnings (Rothfusz et al. 2018). Finally, participants always completed a pre-experiment survey that evaluated their workflow and weather information use, and a post-experiment survey that provided an opportunity for participants to reflect on all of the experimental products and the potential utility in their workflow.

Many of the other aspects about the end user experiment are more fluid based on who is funding the experiment, what the goals of that work are, and how much funding is available. In the spring 2018 experiment, participants completed two simulated events per day. Each one started with a brief (15 minute or so) overview of the weather and then continued directly into the warning exercise (table 3.1). The 2019 experiments shifted the focus to longer-range forecast products. Participants saw SPC products from day 4 down to day 1, watches, and mesoscale discussions before lunch, and then were placed into the warning exercise after lunch (table 3.2). Therefore, the number of cases that were tested were cut in half, but the feedback collected from each case doubled. This shift to the longer range products reflected observations from previous years, when participants were unable to provide authentic feedback because their jobs dictate that they make decisions about resources and staffing well before the warning timescale.

Table 3.2: Example of 2019 EM experiment schedule

Time	**Activity**
9:00a	Tabletop exercise: Convective outlook day 4, 3, 2a, and 2b
9:30a	Tabletop exercise: Convective outlook day 1a, 1b, 1c, and PST
10:00a	Break
10:15a	Tabletop exercise: Mesoscale discussion, tornado watch, and convective outlook day 1d
11:00a	Tabletop exercise: Warn-on-Forecast products
12:00p	Lunch
1:30p	Warning/probabilistic hazard information functional exercise
3:00p	Break
3:15p	End of day survey
3:30p	Group debrief
4:00p	Adjourn

3.3 Initial development

After preliminary research showed that a majority of daily severe weather reports at a single point are contained within a 4-hour period (Krocak and Brooks 2020), researchers developed two different visualizations to display timing information with the probabilistic information currently available in the SPC convective outlook.

The first option consisted of contours of time. These lines represented the starting time of the 4-hour period that forecasters believed the most severe reports would occur within. We wanted to test whether this visualization was difficult to produce in the context of the other forecasting duties. An example of the timing contours, named isochrones, are shown in figure 3.1.

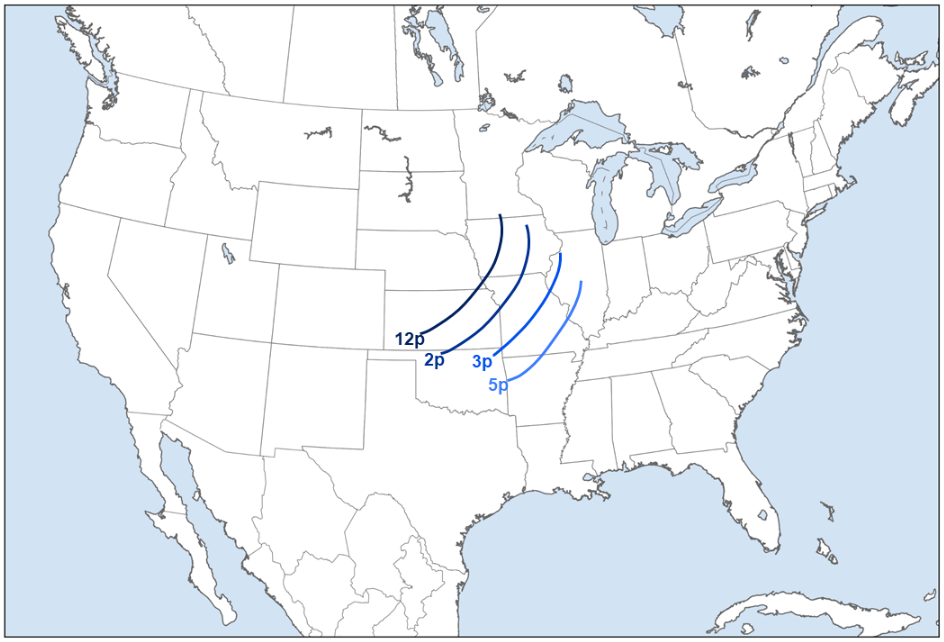


Figure 3.1: One of the original visualizations for a timing information product featuring lines that represent the start time of the 4-hour period forecasted to contain the majority of daily local storm reports.

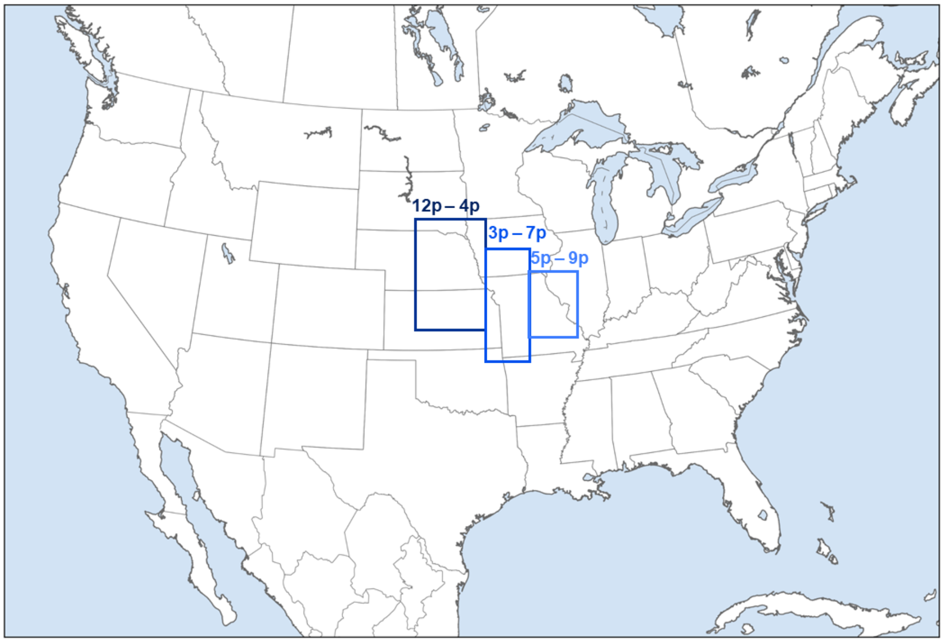


Figure 3.2: One of the original visualizations for a timing information product featuring boxes that represent the time frame of the 4-hour period forecasted to contain the majority of daily local storm reports.

The second visualization option consisted of polygons that represent different 4-hour periods (figure 3.2). Although this option may be more difficult to verify objectively, forecasters would likely be more familiar with the concept since many forecasts currently manifest as some sort of polygon (e.g. watches, warnings, etc.). Ideally, this visualization would resemble the current severe weather watches, following the unwritten ”rules” for watches in size, placement, number of boxes, etc. Additionally, these boxes would not overlap such that each location was within a single 4-hour period. As experiments progressed, it became clear that this was sufficiently difficult and would not be attainable in many situations, meaning locations were often in multiple 4-hour periods on complex forecast days.

3.4 The first iteration: isochrones in the 2016 and 2017 textbed experiments

3.4.1 Data collection

Forecasters in the 2016 and 2017 SFEs were asked to create a completely new product; isochrones. These lines (Figure 3.1) represent the start time of the 4-hour period that forecasters believe will contain the majority of severe weather reports that occur that day. The 2016 SFE occurred from 2 May until 3 June of 2016 with 82 participants representing two different countries and multiple states. The 2017 experiment occurred from 1 May until 2 June 2017 and had 73 participants representing similar geographic areas as the 2016 SFE. It is important to note that not all participants are forecasters, some are researchers or model developers, meaning that explicit forecasting experience is highly variable. Notable cases from these experiments include a large hail event in the Mid-Atlantic on 2 May 2016, a southern plains tornado event on 18 May 2017, and a huge wind event on 27 May 2017 across Missouri, Kentucky, Tennessee, Mississippi, Alabama, Georgia, and North Carolina.

There was very little in the way of isochrone training during the 2016 experiment. Participants were shown a brief presentation about the concept and the goal of the product and then set to work. During 2017, training was expanded to include an hourly area drawing activity to help facilitate the visualization of the isochrones. First, participants were asked to outline the area where they expect severe weather could occur at each hour of the forecast. Then, using those hourly areas, they were able to draw isochrones at more reasonable intervals (instead of one at each hour). This process took more time but allowed participants to get more invested in the process. See 3.3 for a schematic of the isochrone activities in 2016 and 2017.

Since the isochrones were envisioned as an addition to the convective outlook, participants on the innovation desk (representing projects that are still well within the research phase) were asked to draw the probabilistic outlook for any severe weather and *then* begin drawing the isochrones. This process was done from roughly 1500 UTC until 1630 UTC

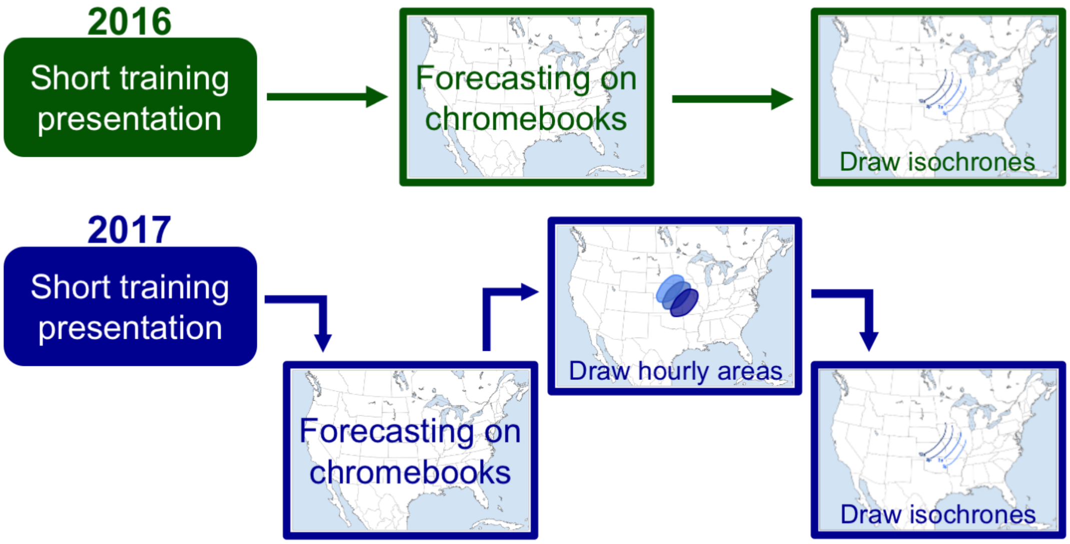

2016
Short training presentation
Forecasting on chromebooks
Draw isochrones
2017
Short training presentation
Forecasting on chromebooks
Draw hourly areas
Draw isochrones

Figure 3.3: A schematic of isochrone activities in the 2016 and 2017 SFE.

each day. First, the group would collectively draw the probabilistic outlook, then individuals would draw the isochrones on chromebook computers separately. Forecasters had access to whatever guidance or observational tools they wanted to use during both portions of the activity. The chromebooks were connected to the internet to allow participants to access whatever they deemed useful without restrictions.

The day after forecasting the isochrones, participants were shown their forecast alongside local storm reports that were aggregated into 4-hour time periods. I also developed an ”automatic isochrone” method to help visualize the *best* forecast. Reports were gridded onto running 4-hour grids (such that any single report could be plotted on multiple grids) that were 80km in horizontal resolution, and then smoothed with a 120km Gaussian kernel. This created smoothed verification grids at running 4-hour intervals, which were then compared to each other. Each point was compared individually. The time period of the grid with the highest value was chosen for each point and then those time periods were contoured. This method resulted in maps that resembled isochrones but could become quite messy on days with multiple rounds of severe weather or if a system was moving slowly.

Researchers relied on observation and survey instruments to collect data about forecaster interpretation and challenges. Participants were very open about their enthusiasm for the concept and concerns about the implementation. During the 2016 experiment, the only survey data collected was during the daily verification period. During this time, participants were asked to evaluate the previous day's forecast and comment about the challenges associated with creating the forecast. During the 2017 experiment, similar daily evaluations were complimented with a weekly review of the isochrone concept and potential avenues for improvement (see appendix A.1.2).

3.4.2 Lessons learned

The 2016 and 2017 experiments had a strong learning curve for both researchers and participants. It quickly became clear in 2016 that the isochrones were a far more challenging endeavor than originally anticipated. Part of the process of testing new products is evaluating the workload to create the information. We discovered very early in the 2016 experiment that participants needed more time: more time to hear about the concept and background research that shows it is feasible to assign times to severe events; more time to ask questions about the drawing tools and possible forecast philosophies; more time to explore the data and synthesize it into a coherent story. Overall, participants needed more time. They also needed more training and feedback from researchers and forecast verification. After the 2016 experiment, I developed a real-time version of the verification method described above using local storm reports. These ”observed isochrones” were shown in real-time in 2017 so participants could see how their forecast did and calibrate to their personal biases. The verification isochrones are simply the observed start time of the 4-hour time frame minus the forecasted start time. An example of the forecasted isochrones (created by the lead forecaster), the observed isochrones, and the verification are shown in Figure 3.4.

In addition to the difficulty forecasters expressed in creating the isochrones, an objective analysis of the 2016 experiment forecasts showed that participants were placing isochrones

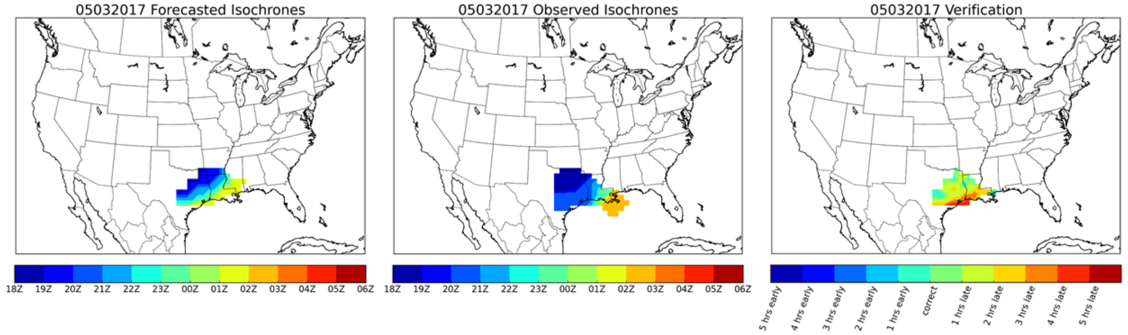


Figure 3.4: Isochrone verification plots showing the forecasted time frames (left), the observed time frames (middle), and the observed minus the forecasted time frames (right).

such that the forecasted propagation of severe storms was generally too slow. Most forecasted time frames during the 2016 experiment were two to four hours later than what they should have been (Figure 3.5). This phenomenon was also referenced in the participants' subjective evaluations of the previous day's forecast:

"We were often too slow moving things east off the dryline."
"The isochrones were too progressive but the initial isochrones captured the western extent of the reports very well."

Given the difficulty creating the isochrones coupled with the inaccurate forecast times, we decided to implement additional training time and an additional forecast activity to help participants synthesize the mountain of guidance information into time periods for each location. In 2017, instead of going straight to contour lines of time, participants first circled areas on the map where they believed severe weather would occur at each hour. Then participants could visualize the movement of the system through both time and space on a single map. The addition of this activity in 2017 greatly reduced the confusion and anxiety felt by participants, and it improved the forecasted time periods as well (Figure 3.6). In the 2017 experiment, most forecasted points were either correct or 2 hours late (an improvement over 2016, when most locations were forecasted to be 2-4 hours late). These issues and subsequent improvements to the forecasting process would not have occurred without the in-depth testing process and feedback gathered from forecasters during the experiment.

Another concern that was raised in the 2016 and 2017 experiments was the difficulty of forecasting time frames for marginal events. There was significant feedback in the discussions and survey data that showed participants were frustrated when they had to identify time frames for weakly forced events because the unknown of whether or not the event is even going to happen makes forecasting the timing of the event nearly impossible. After the 2017 experiment, we decided to limit the forecasting of any timing information to days and areas with at least a 15% probability to reduce the frustration with the process and the selection of an arbitrary time frame to complete the activity.

”I am not sure the isochrones are useful for such a marginal event.”

”[The forecast was] very difficult to do with few reports.”

”It is hard to rate with 10% or less [probabilities] and very sparse reports.”

”Based on the low number of reports (3 hail reports), it is hard to have a strongly positive or negative opinion on this forecast.”

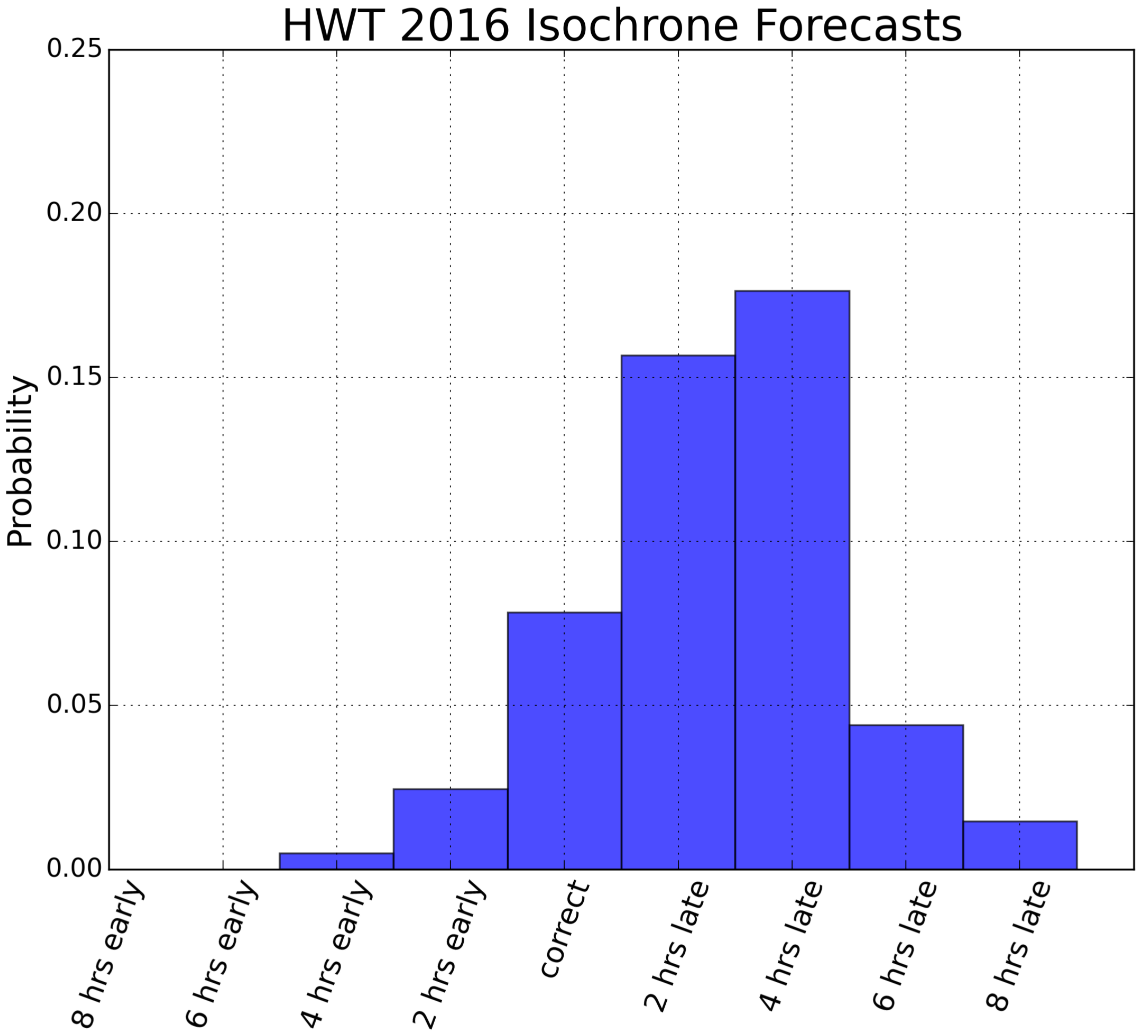


Figure 3.5: A histogram of the aggregated isochrone verification shown in Figure 3.4 for the 2016 SFE.

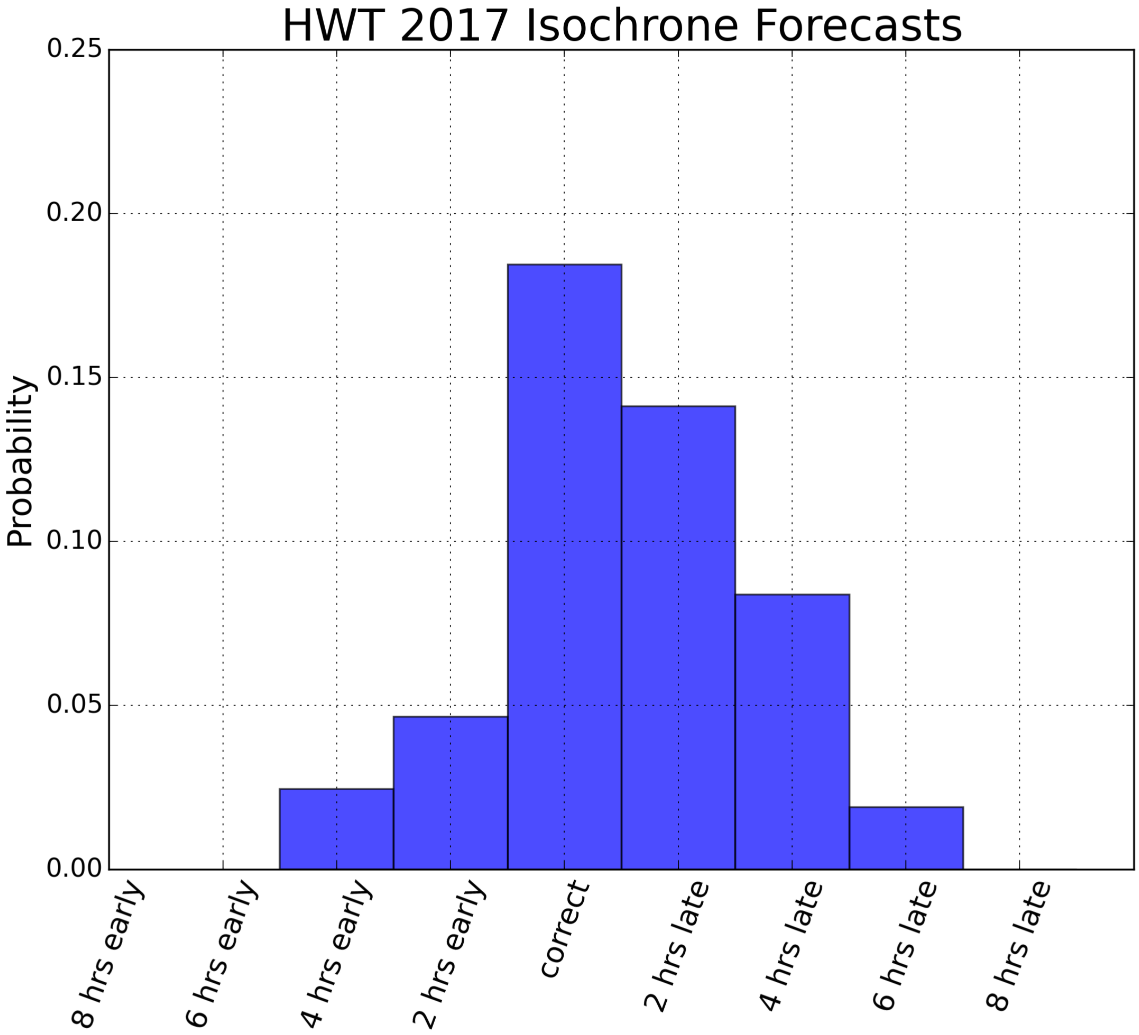


Figure 3.6: A histogram of the aggregated isochrone verification shown in Figure 3.4 for the 2017 SFE.

Finally, the biggest lessons learned during the 2016 and 2017 experiments were related to the interpretation of timing contours. Most products that are created for severe convective weather are visualized as areas (i.e., the convective outlook, the weather watch, the warning), therefore it is difficult for forecasters to create a product that is easy to interpret for multiple audiences with lines of time. Even on more straightforward days, forecasters expressed concern that the lines of time would be interpreted as the entire duration, instead of the start time of a 4-hour period. For example, if a 00 UTC line was west of a 02 UTC line, participants expressed concern that people would interpret the threat time as 00 to 02 UTC, instead of 00 to 04 UTC.

”The problem with conveying this information in an operational environment to the public as well as to the forecasters is the confusion of how peak severe would fall between the two lines.”
”I think there is still confusion. The isochrone is the ”back edge” of the severe reports. We (some of us) were looking at the isochrone as ”the leading edge.” Thus our lines were too far east.”
”[Isochrones are] interesting, but hard to interpret...and not immediately intuitive.”
”The most difficult challenge was knowing when to draw isochrones after the first isochrone. For example, it was relatively easy to draw the 18-22Z isochrone but I did not understand when to draw a 20-00Z isochrone and what that would mean.”
”I think this is the most complicated part. If forecasters are struggling with the methodology used to draw the isochrones, it will be even more difficult to explain it to the public.”
”I don’t think this product is likely to be used properly by the general public, but it may be useful for EMs, event planners, etc. With proper training, the product should be useful as-is for such users.”
”I’m sorry I’m not particularly sure how [isochrones] can be improved. What a tough problem!”
”When there’s only one isochrone and beyond the last isochrone, there is potential ambiguity about when or whether there is a severe threat.”

On more difficult forecast days, the interpretation confusion was compounded when participants felt that they needed to highlight a time frame longer than 4 hours. Since contour lines should never cross, participants have no way to indicate a potentially longer time frame or uncertainty in a particular forecast. This made participants very wary of this product becoming operational as it does not allow the flexibility needed to display complex or uncertain forecasts. Furthermore, in addition to raising these concerns, participants could not envision a way to reduce this confusion with the contour line visualization.

”The biggest limitation with this particular forecast was multiple areas of convection – do you choose the most active area and assign isochrones there or attempt to time 2 or 3 separate areas?”
”Isochrones for this case were very tough, as there we multiple rounds of severe convection tracking over the same area. The isochrones are about as good as could be expected for this case.”
”It’s difficult to understand exactly what [isochrones] are supposed to show, especially in situations with complex evolution, back-building, and multiple rounds in the same area.”
”We need to gain clarity on how to ’derive’ the isochrone position from partially overlapping hourly areas.”
”Sometimes, the greatest severe threat extended over slightly more than four hours within a relatively confined location. It was difficult to know how to position an appropriate isochrone and what time to ascribe to it in such scenarios.”

Ultimately, the 2016 and 2017 experiments exposed researchers to just some of the operational challenges that forecasters face when predicting severe weather events. Even with proper training and experience, we were not convinced after the 2017 experiment that the isochrone visualization was a worthy forecast product to continue to evaluate. Luckily, we learned many valuable lessons in the two years of isochrone experiments that paved the path for a new visualization to test: timing *areas*. The most influential lessons learned in the 2016 and 2017 experiments are summarized in Table 3.3.

Table 3.3: Lessons learned in the 2016 and 2017 SFE experiments and future actions to consider.

Lesson	**Future steps**
The drawing process is confusing and cumbersome	Longer, more in-depth training activity and a more detailed drawing process (including hourly report areas).
Verifying isochrones with just local storm reports is inadequate	Showing the verification method daily to help forecasters calibrate.
Isochrones are consistently forecasted too late	Have participants draw hourly report areas to synthesize the guidance information and their thoughts on one map.
It is very difficult to forecast time frames for marginal events	Only require timing forecasts for areas with a 15% probability or higher total severe forecast.
Isochrones are time intensive to create and there are inconsistent, conflicting interpretations of the forecasted time frame	Try a new visualization: timing *areas*.

3.5 The second iteration: PSTs in the 2018 testbed experiments

3.5.1 Data collection

After two years of testing isochrones with forecasters, it was clear that this visualization was not going to be ideal for daily creation. Therefore, the 2018 experiments ushered in a new timing visualization technique: the Potential Severe Timing (PST) product. This product, which utilizes boxes instead of lines (Figure 3.2) presented a different set of challenges than the isochrones. Areas are more difficult to verify objectively, which may lead to forecaster calibration issues. However, after testing the lines concept (Figure 3.1) for multiple years and seeing just how difficult creation and interpretation was, the potential calibration issue seemed like a reasonable trade-off.

The 2018 SFE occurred from 30 April 2018 to 1 June 2018. The biggest change that occurred during this experiment was the change in the lead forecaster. Since the product being created was also changing in 2018, it is difficult to extrapolate what changes in creation and interpretation were due to the change in the product versus the change in the lead forecaster.

Similar to previous years, the PST was tested on the innovation desk during the morning and early afternoon activities. After drawing the 16 UTC to 12 UTC probabilistic forecast, participants were asked to draw areas that indicate the 4-hour period that would contain the majority of severe weather reports at that location. Then after lunch, participants updated these timing forecasts with the latest guidance information. These forecasts were created in small groups using individual machines and the new forecast drawing tool developed by the Cooperative Institute for Mesoscale Meteorological Studies. This tool was a vast improvement over previous years because it is web-based, intuitive, and similar to other websites that show model guidance information. When forecasting these areas, participant groups were limited to specific model subsets so researchers could illicit information about the quality of individual model configurations.

There were 86 participants from two different countries and multiple US states during the 2018 experiment, showing a diverse set of skills and perspectives. Notable cases from the 2018 experiment include a classic tornado and severe weather event on 01 May in Kansas, Nebraska, and Iowa (Figure 3.7, top); a challenging wind event on 14 May in Missouri and Illinois (Figure 3.7, middle); and an overnight wind event on 30 May in northern Oklahoma (Figure 3.7, bottom).

Before participants began drawing the first set of PST areas each week, researchers presented a brief training presentation that included a discussion of the PST concept (including

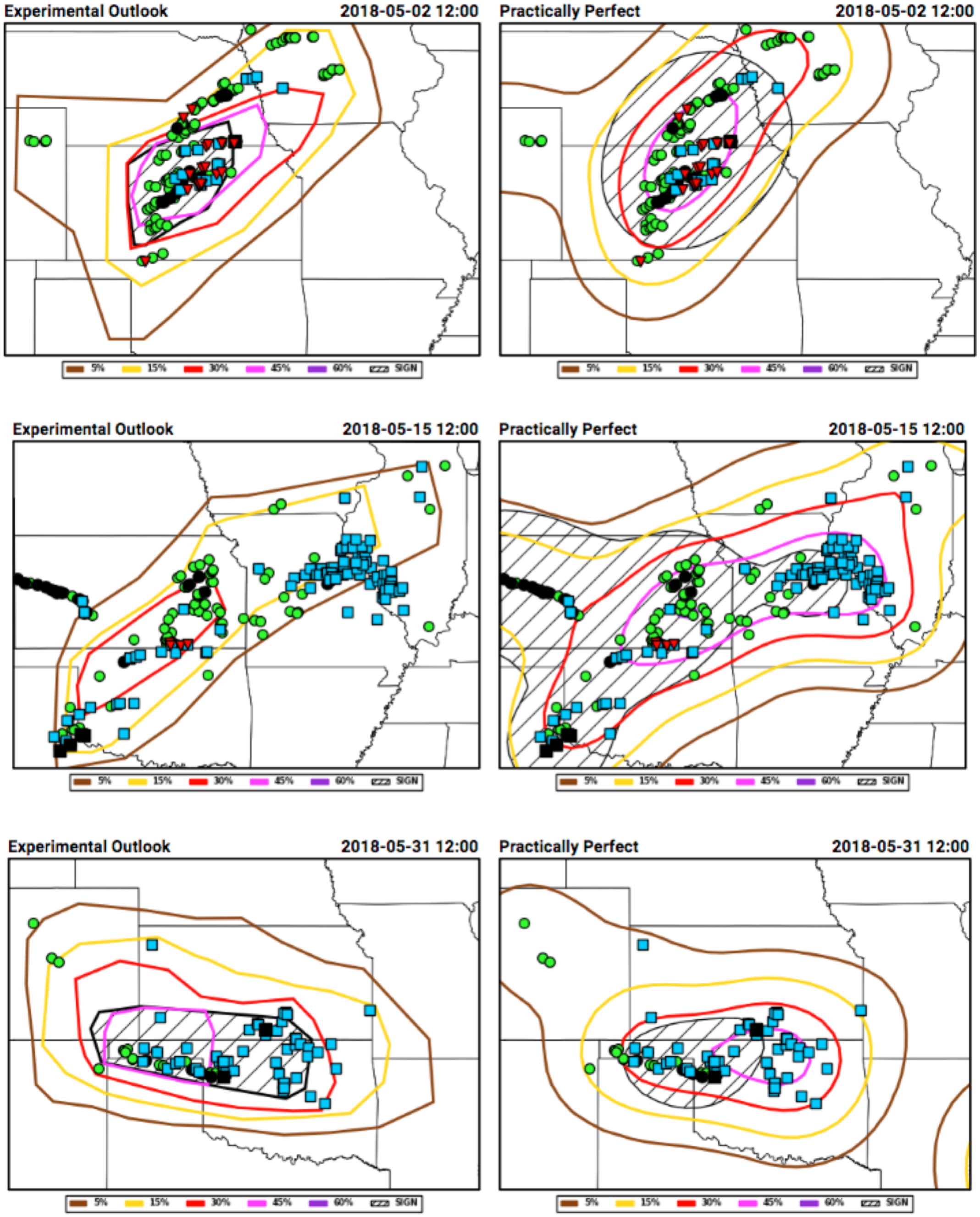


Figure 3.7: Examples of of the probabilistic total severe weather forecasts (left) and the practically perfect probabilistic forecasts (right) for notable cases in the 2018 SFE. Green dots are hail reports (greater than 1 inch reports are black dots), blue squares are wind reports (over 65 knots are black squares), and red triangles are tornado reports (EF2 and greater reports are black triangles).

past research that showed it was theoretically possible to capture a majority of daily reports in a four hour period) and the goals of the visualization. Then forecasters were instructed to find a small group and begin forecasting in the drawing tool. The morning after the participants created the timing forecast, local storm reports were displayed hourly on top of the PST areas to attempt to verify the forecast. Participants were instructed to rate the forecast on a 1-10 scale and then answer questions about the difficulty of creating the PST, the utility of the product and the biggest challenges (see appendix A.1.1 for sample questions).

Finally, another new activity in the 2018 experiment was a forecaster discussion period directly after participants finished drawing. This activity was initially driven by participants who wanted to hear about what others were seeing in their guidance subset. After a few of these discussion periods, researchers realized this activity was incredibly valuable because forecasters were comparing and contrasting guidance information *and* PST forecasting philosophy. Participants learned from each other and researchers could more holistically discuss different forecasting techniques.

The biggest addition to the timing information development was the inclusion of end users in the 2018 experiments. Stakeholders had never before seen a prototype timing product or provided input on the utility of such information. While forecaster participants and researchers had long speculated that this information would be useful for emergency managers, this group of people had never actually been involved in the discussion.

The 2018 end user experiments were largely focused on prototype warning-scale products, including probabilistic hazard information (Miran et al. 2018a). Given that this was the focus of the experiment and not longer lead time products, researchers were grateful to have the opportunity to present the PST to the participants. During the simulated events, participants were shown the day 1 SPC convective outlook and asked about the decisions they would make. Then, participants were shown the PST and asked if any of the decisions or time frames for decisions had changed before continuing on to the warning-scale exercise. After the simulated case was complete, participants filled out a survey that asked them if the PST helped them prepare for the warning simulation and what (if any) information was missing. Participants completed two simulated cases per day, one in the morning and one in the afternoon, for a total of six cases completed each week.

There were eight EM participants in the 2018 end user experiment. These participants were selected based on diversity of jurisdiction type, geography, and knowledge of the current weather warning system. Participants ranged in geography from Minnesota to Texas, and from jurisdiction type from cities to states and healthcare systems. A table summarizing the 2018 participants can be found in appendix B.1.

The six simulated cases in the 2018 experiment represent a wide variety of weather and geography. The cases include a couple Kansas tornado cases (25 May 2017 and 25 May 2016), a wind event near Jackson, Mississippi on 21 January 2016, a multi-hazard case in South Carolina on 24 May 2017, a lightning case in Grand Junction, Colorado on 22 July 2016, and a labor day case in Florida on 1 September 2016. The simulated PSTs that were created for these cases were purposely simplistic (Figure 3.8). They were formulated to look similar to current watches, except they included a forecasted timeframe that was later in the day (Figure 3.8).

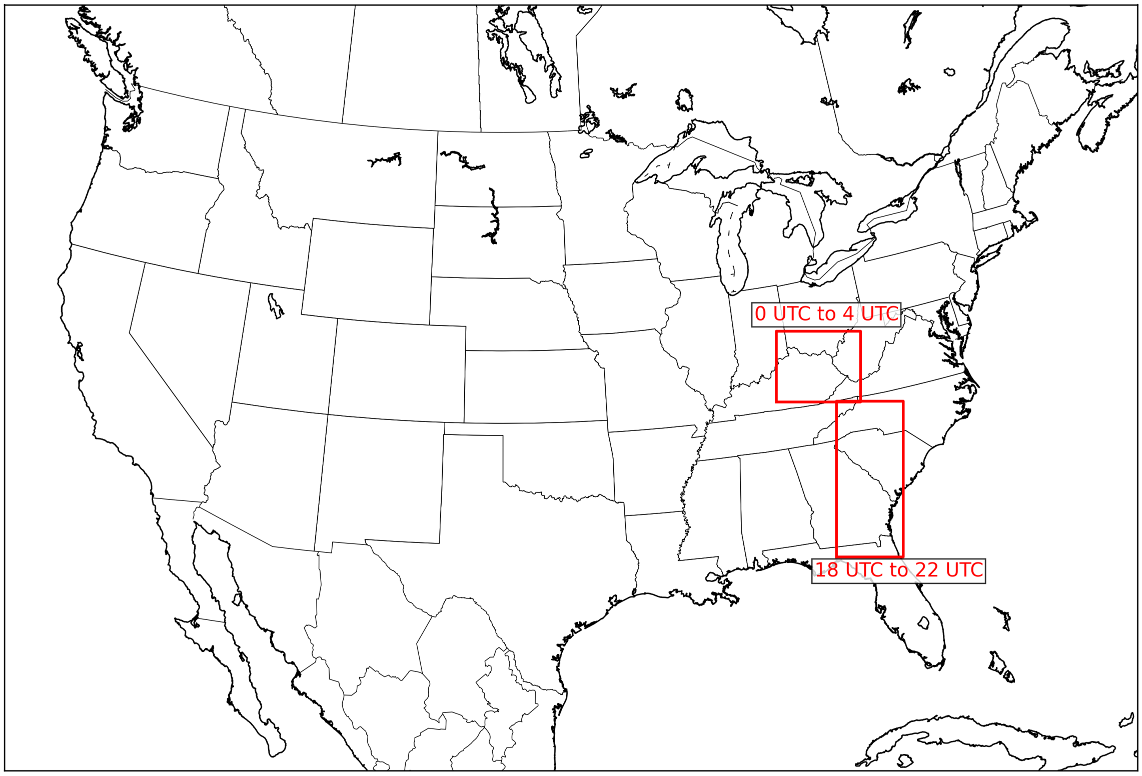


Figure 3.8: An example of the simulated PST product shown to EMs during the 2018 experiment.

3.5.2 Lessons learned

The 2018 SFE provided an opportunity for researchers to test the second timing visualization in a very similar environment to the isochrone testing that took place the two years prior. Many of the participants that evaluated the PST were also involved in testing the isochrones, which gave some sense of longitudinal data and evaluation. First and foremost, nearly all of the participants who interacted with both visualizations agreed that the PST (or areal) visualization was easier to create and tell the timing story than the isochrone (or contour) visualization. Particular situations especially lend themselves to an areal approach, like very slow moving systems or days with potentially multiple rounds of severe weather. Overlapping or crossing of lines does not make sense in a single graphic, but overlapping polygons allow forecasters more freedom to express uncertainty or longer time periods.

”I could see this being really useful operationally to provide greater information on the timing windows.”

”It does seem like this could be a useful tool, and suggests it may be possible to issue ’watches’ well ahead of time.”

”Great idea - these will be extremely valuable to emergency management and the public. I’m surprised these hadn’t been put in place sooner, and using the different ensembles is a unique way to test it.”

”I think the PST areas are conceptually easier to understand than isochrones.”

Overall, the PST concept was much more widely accepted than the isochrone visualization, both from a creation and an interpretation standpoint. During the drawing process, researchers saw different forecasting philosophies manifest in wildly different PST areas. Some forecasters were very methodical in choosing a single 4-hour timeframe for each point (Figure 3.9), while others were more concerned with covering the entire timeframe that severe reports were possible (Figure 3.10). These different philosophies were much more pronounced than researchers had anticipated. The discussion portion of the activity largely focused on discussing the reasoning for these different philosophies and what the different portrayals of timing may indicate to end users. Forecasters were mostly concerned with the lack of consistency and clarity driving end users to look to other sources for timing information. However, they did not have a clear solution because they were grappling with needing the flexibility to draw overlapping areas when the events were less certain, but also wanting to provide a clear and easy to interpret story for those using the product to make decisions. Researchers quickly realized that best practices would need to be developed to aid forecasters in deciding how to draw the areas with different philosophies and an array of meteorological challenges.

Similar to the challenges with training and best practices, forecasters also found the verification process to be more complicated than expected. During the morning *verification of yesterday’s forecast* period, participants were shown the previous day’s PST areas and

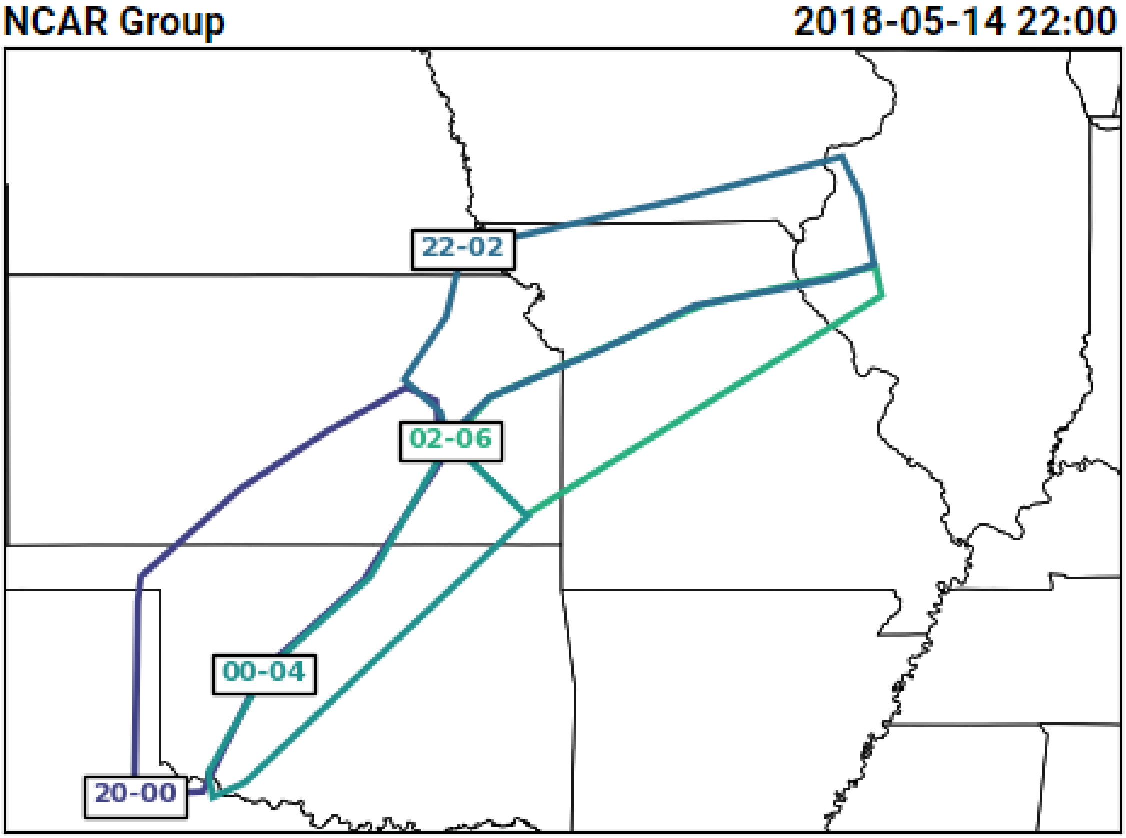


Figure 3.9: An example of forecasted PST areas drawn by participants during the 2018 SFE.

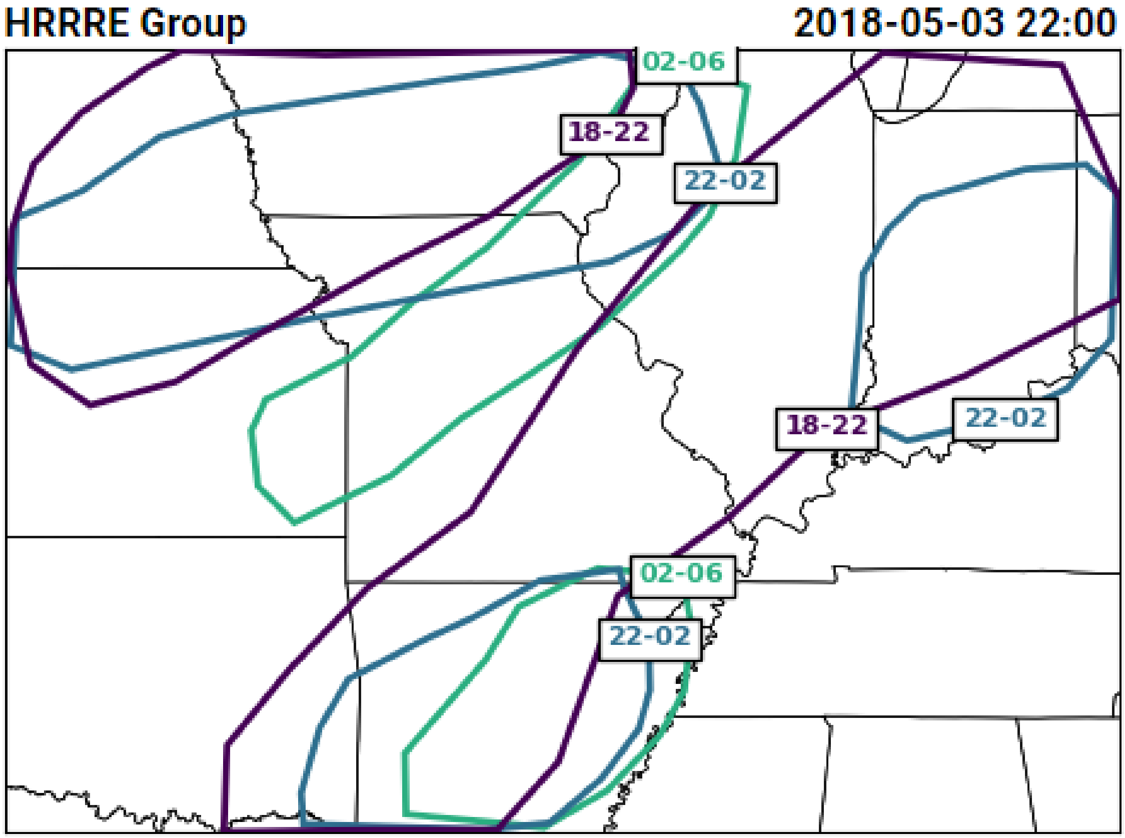


Figure 3.10: An example of forecasted PST areas drawn by participants during the 2018 SFE.

then an hourly rotation of the local storm reports. However, since PSTs cover both time and space, a single report can verify multiple PST combinations. Therefore, participants had a difficult time evaluating the quality of the PSTs with just individual reports. What was the *best* set of PSTs? How much *worse* were all the rest? These are important challenges that were uncovered during the 2018 testing process. Forecasters cannot be expected to create a product on a daily basis without objective feedback about the quality of the forecast. Users also cannot be expected to blindly trust a product without confirmation that it is high quality information. Future work needed to address this issue by developing a concept for objectively evaluating the PSTs in quasi-real time, similar to the method that was developed for the isochrones.

In addition to the challenges associated with understanding the quality of the PST forecasts, there also is not a comparable product to evaluate the PST against. One may argue that weather watches are similar in spatiotemporal scales, so researchers started by comparing PSTs to them. The performance diagram in Figure 3.11 shows that PSTs drawn during the 2018 SFE had a higher probability of detection (the number of hits divided by the sum of hits and misses) than the watches issued during the same time periods (0.79 vs. 0.64), but also a lower success ratio (1 - the number of false alarms divided by the sum of hits false alarms, 0.23 vs. 0.37). This is likely due to size and placement differences between the two products. The PSTs are generally forecasted earlier relative to the reports than the watches are, meaning forecasters are less certain about the exact area and extent that reports will occur within. Therefore, PSTs are larger in area, capturing more of the reports, but also leading to more false alarm area. Regardless of these differences, it was promising to see the PSTs show some skill with respect to an operational product with fewer temporal restrictions.

One of the major developments in the 2018 experiments was the initial discussion of overlapping PSTs. When forecasters were using isochrones, they felt that they couldn't overlap lines since that created all kinds of interpretation issues. However, the polygon

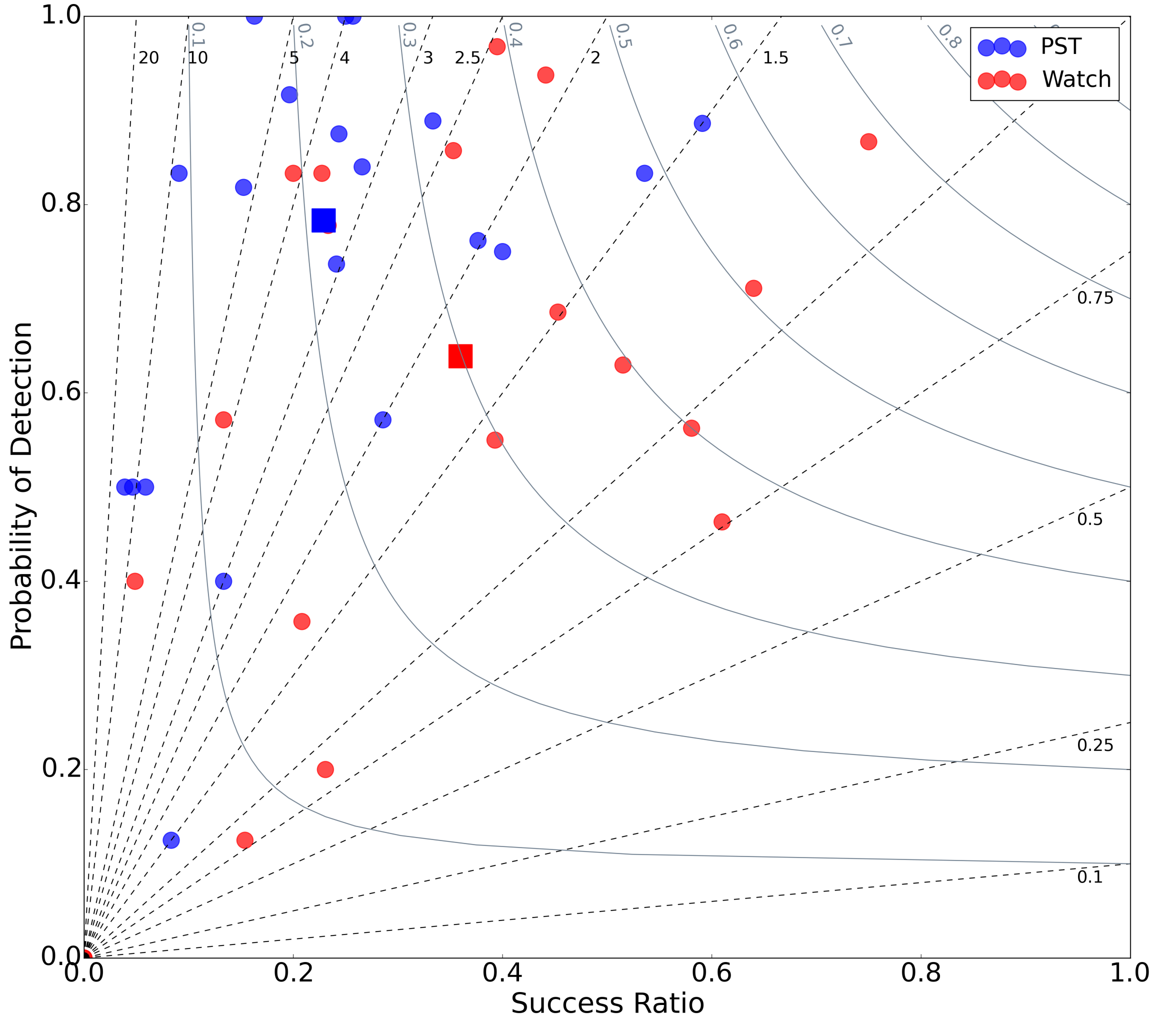


Figure 3.11: A performance diagram showing the probability of detection (y-axis) and the success ratio (x-axis) of the 2018 SFE PSTs (blue) and operational weather watches issued during the same time period (red). Each dot represents a single day, and the larger square is the mean performance.

visualization was much more straightforward, making forecasters more comfortable with overlapping. When researchers discussed this overlapping area with participants, there was no clear consensus on what it meant. For example, one participant may overlap a 00 UTC to 04 UTC area with a 04 UTC to 08 UTC area and interpret the overlap to indicate a 00 UTC to 08 UTC timeframe (i.e. the entire period). Another participant may interpret the overlap area to indicate a 02 UTC to 06 UTC time frame (i.e. the middle of the time period). This example only included spatial overlap, but there were many participants that actually included both spatial and temporal overlap, compounding the interpretation issues. During the discussion periods, participants also brought up potential confusion with end user interpretation. If forecasters could not agree on what information they were creating, how would end users be able to? Future experiments required that researchers more holistically explore this philosophy and possible solutions or compromises about if, when, and where overlapping areas are necessary.

The EMs in the 2018 end user experiment were exposed to two completely new forms of communicating weather information. The PST was presented during the the simulated weather briefings and during a live weather discussion. The overall concept of providing the time frame of severe weather was well-received by participants. Nearly all of the participants agreed that this information would be useful to provide awareness of when the event would be most threatening. Depending on the situation, the timing information would also be useful to help aid in staffing and resource decisions, including school closures. Overall, participants noted that there were certain time frames that were more concerning for their operations, and those decision points are what they look for in the SPC convective outlook discussions.

"If there's anything the day shift can do to help prepare, they'll do it. If [the event] is more towards 11 PM, I would be thinking about the night shift not making it in." - EM 1

"Since it is the end of the school year, I would be looking at field trips. The time of the year means schools will be high on our notification list. There are also events like track and field, outdoor events, etc." - EM 4

"This is a great product, I look for MCDs for timing but this is much easier." - EM 2

"If I am in a slight risk or above, I'll go look at ensemble models to get timing info. I usually make my own timing graphic to Tweet, but this seems like it would be a lot easier and more generalized." - EM 8

Although the EMs gave overall positive feedback about the PST, they did note that they would have to spend more time than what was allocated investigating the quality of the product and gaining confidence that it provided accurate information. Some participants even said that without the convective outlook, the PST did not make much intuitive sense. Researchers concluded that future experiments need to include the convective outlook and the PST as complimentary products, and also include more time to evaluate the utility of the PST.

Similar to the previous point, participants expressed concerns about knowing when and where to look for such a timing product. Six of the eight participants reported using the SPC website to look for updated information, but noted that others do not have the luxury of time to go searching. Participants largely agreed that this information is very important and needs to be readily available in a easy to access location.

Other smaller points that the participants expressed were concerns with colors and with timezone labels. While most forecasting products are displayed in UTC time, this isn't useful for EMs. This issue was again related to a lack of time to interpret information. If they have to do manual conversions to local time zones, they are unlikely to use the product. However, they also sympathized with the local timezone issue and mapping information. If a PST area spanned multiple time zones, which one should the label be displayed in? As for colors, what do they mean? Is red worse than blue? Does it mean more storms? Earlier times? More intense storms? While these issues may seem small, they will all contribute to the utility of the product and whether or not it gets used.

Finally, EM participants expressed similar concerns to the forecasters regarding overlapping PSTs. They had a wide variety of interpretation and acceptance, with some thinking the overlap told a coherent story about the evolution of the storms, and others expressing concern about what they would do if they were in the overlap area. Some participants thought that the overlap meant there would be two rounds of storms, others thought that the overlap was where the worst weather would be expected (because the overlap makes that area look like a bullseye). And then others thought that the overlap was actually where forecasters were *most* confident in the timing of storms. Given the little conversation that was devoted to overlapping PSTs in the 2018 experiments, further work should include a formal evaluation of what the interpretation of overlapping areas is and if/when they are necessary.

The 2018 HWT experiments introduced forecasters and EMs to a new concept: the potential severe timing product. These experiments were largely focused on evaluating the difficulty in creating the product and the potential utility for users. Few details were thoroughly explored, but some overarching challenges were uncovered. Future experiments should devote more time to clarifying the forecasting process and a more detailed evaluation of the utility for end users. Specific lessons and potential changes or improvements are listed in table 3.4.

Table 3.4: Lessons learned in the 2018 experiments and future actions to consider.

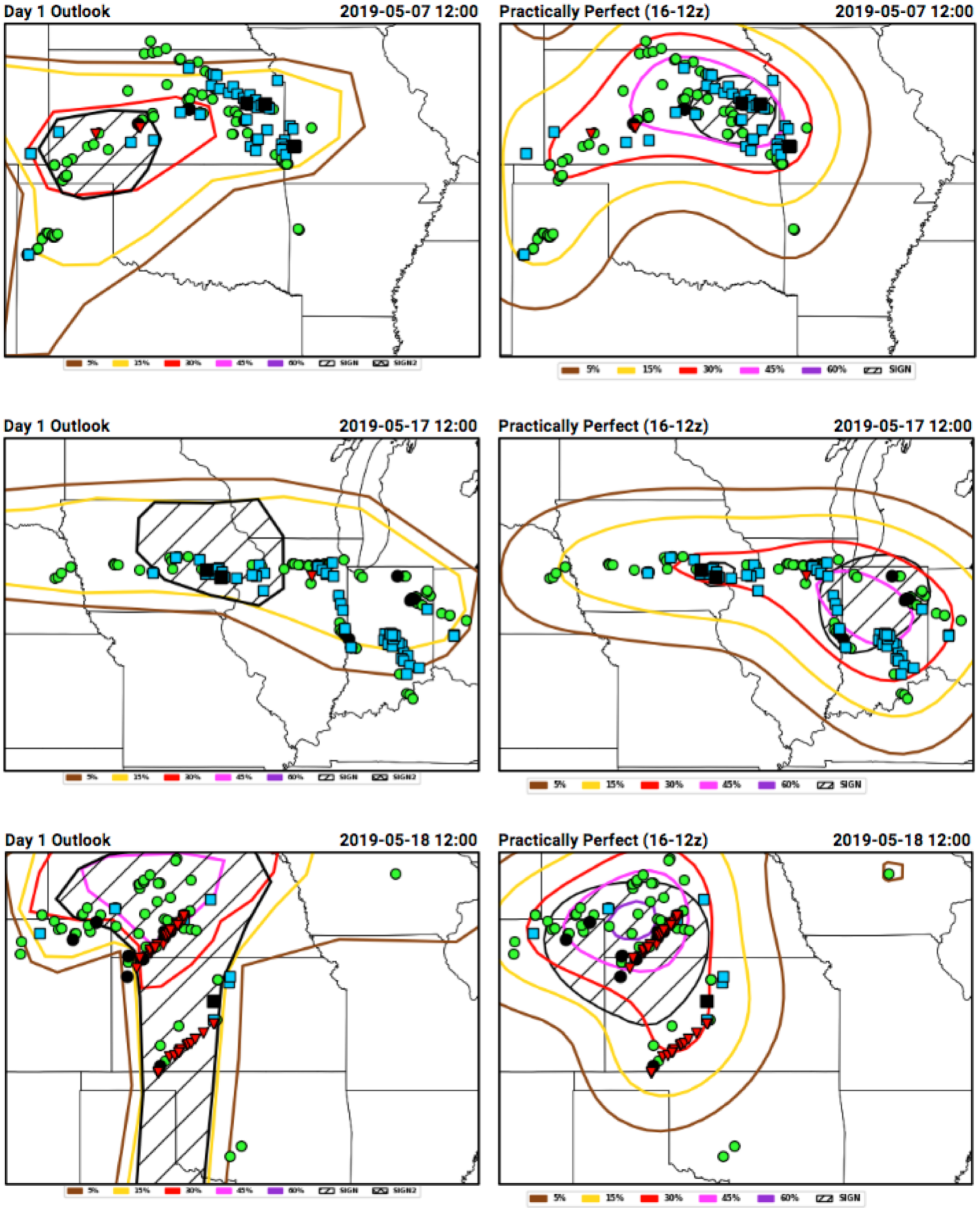
Lesson	**Future steps**
Forecast philosophies are highly variable	A more thorough training procedure and the development of best practices would be helpful.
Verifying PSTs with single reports is confusing	Developing an objective verification method is critical.
Overlapping areas are confusing and understudied	A formal evaluation of overlap interpretation and necessity should be prioritized in future experiments.
The PST info alone is incomplete	The product should be presented in tandem with the SPC convective outlooks.
UTC is not well understood by users	Prototype graphics for users need to be presented in local time.
Color has many interpretations	PSTs should be displayed monochromatically.

3.6 Continued refinements: the 2019 testbed experiments

3.6.1 Data collection

The 2019 experiments ushered in a few changes on the forecast side and significant changes in the end user experiment. The SFE ran from 29 April 2019 to 31 May 2019 and included 94 participants from three countries who participated in the 8am to 4pm regularly-scheduled activities. Notable forecasting cases include a wind and hail event on 6 May in Texas (Figure 3.12, top), a messy mesoscale convective event in Iowa, Illinois, and Indiana on 16 May (Figure 3.12, middle), and a few long-lived tornadoes on 17 May in Kansas and Nebraska (Figure 3.12, bottom).

There was once again a new lead forecaster on the total severe desk. The morning began with an evaluation of yesterdays forecast, then map analysis and the creation of



Day 1 Outlook 2019-05-07 12:00

Practically Perfect (16-12z) 2019-05-07 12:00

Day 1 Outlook 2019-05-17 12:00

Practically Perfect (16-12z) 2019-05-17 12:00

Day 1 Outlook 2019-05-18 12:00

Practically Perfect (16-12z) 2019-05-18 12:00

Figure 3.12: Examples of of the probabilistic total severe weather forecasts (left) and the practically perfect probabilistic forecasts (right) for notable cases in the 2019 SFE. Green dots are hail reports (greater than 1 inch reports are black dots), blue squares are wind reports (over 65 knots are black squares), and red triangles are tornado reports (EF2 and greater reports are black triangles).

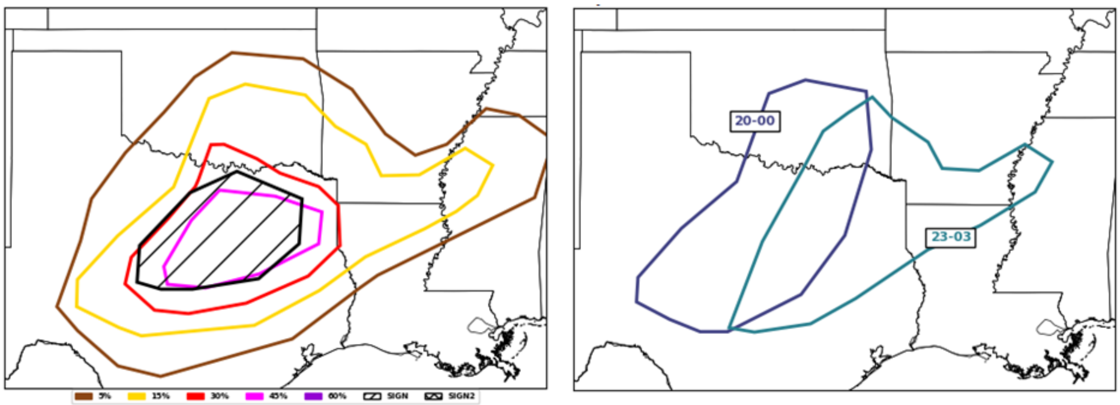


Figure 3.13: An example of of the probabilistic total severe weather forecast (left) and the PST timing areas (right) that were drawn in the 2019 SFE.

the probabilistic total severe weather forecast (Clark et al. 2019). On Monday of each week, participants were given a brief training exercise to introduce them to the concept of the PSTs and some best practices for creating the product that were developed over the previous testing periods. These best practices include i) cover the entire 15% area, ii) don't draw an area for every hour, use only the time periods that will have severe weather reports occur, iii) minimize overlap, and iv) keep it simple. Many of these best practices were developed after forecasters the previous year had difficulty with the concept. This training session was followed by questions and discussions to further clarify any difficulties before participants started drawing.

After the probabilistic graphic was created together as a group (Figure 3.13 left), forecasters set out on the task of creating the PST areas individually (Figure 3.13 right). Forecasters created these timing areas on individual machines using unique ensemble model subsets so researchers could identify strengths and weaknesses of different model configurations.

After forecasters created the PST areas, researchers led a 10-20 minute discussion about the forecast, including challenges, what the different subsets were keying in on, whether or not forecasters trusted the model guidance, and overall philosophy when drawing the timing areas. This discussion was often one of the most fruitful activities of the experiment, as it allowed forecasters to openly discuss what did/did not go well during forecasting process they just finished. It also allowed participants to hear other about ideas and philosophies, often surprising each other (and researchers) with new methods for evaluating guidance and synthesizing the information into the PST graphic.

The morning after the forecasts were created, participants would evaluate the quality of the PST areas based on report locations. New to the 2019 experiment was an objective verification technique called *automatic PSTs*. This method was introduced in 2019 after feedback in the 2018 experiments indicated that it was difficult to evaluate the quality of PST areas by reports alone. Since single reports can verify multiple PST areas (given that they cover 4-hour time frames), there are multiple combinations of PSTs that would be viable. Therefore, the automatic PSTs were developed as a method to create the *best* combinations of PST areas. The algorithm works by plotting local storm reports on 80 km spatial grids that are valid for running 4-hour periods starting at 18 UTC and ending at 2 UTC (i.e. 18–22 UTC, 19–23 UTC, 20–00 UTC, ... 02–06 UTC), which means that reports were plotted on multiple grids. After plotting, the grids were smoothed using a Gaussian kernel with a 120 km smoothing parameter. Then a threshold of 15% of the maximum value was applied to the smoothed grids to create areas for every 4-hour period. This threshold was chosen for a couple of reasons. First, it captures a majority of the reports on a day. Figure 3.14 shows the PDF of daily POD for the forecasts created by the algorithm for the 2018 HWT cases. When the threshold was set at 10% of the maximum grid value, the POD for the cases peaked at about 0.9. With a threshold of 15% of the maximum grid value, the peak POD was slightly lower. A threshold of 20% dropped the peak to about 0.75. Next, the PDF of the area under a PST created by the algorithm (Figure 3.15) shows that the

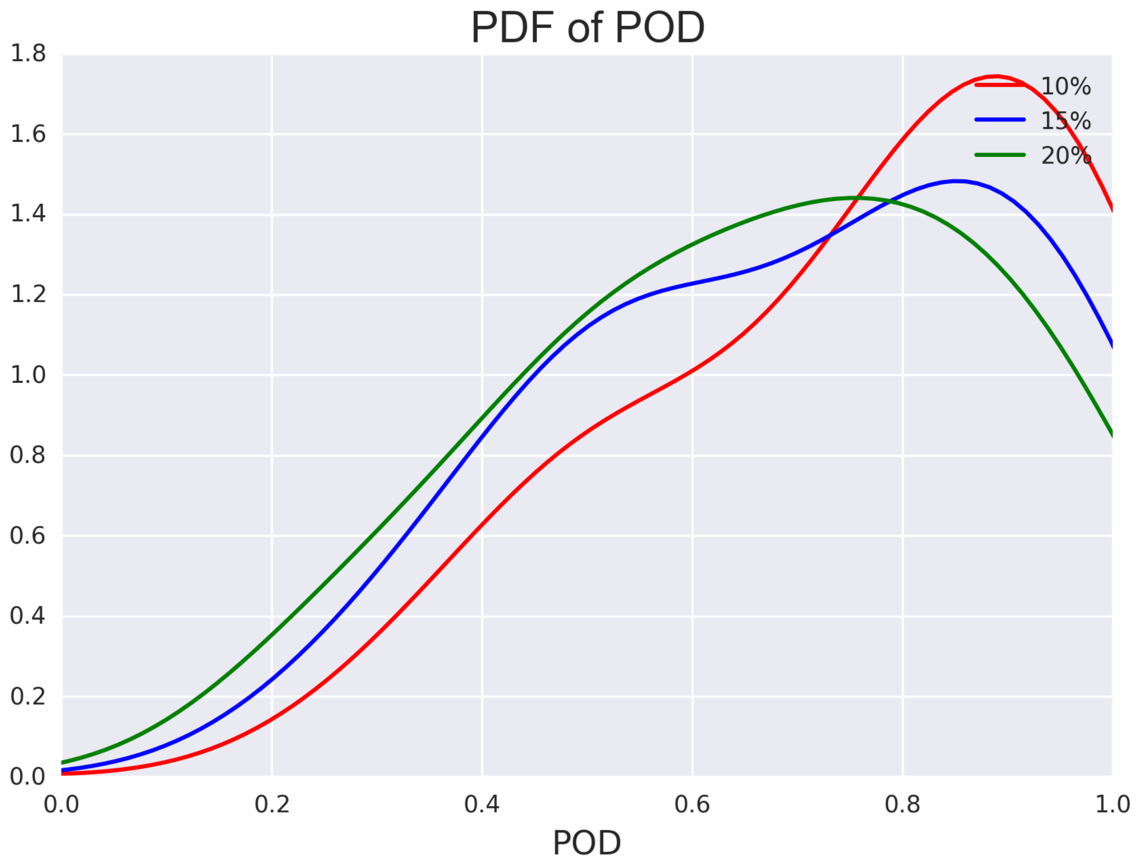


Figure 3.14: The probability density function of the probability of detection for automatic PST areas generated using three different thresholds (noted as a percentage of the maximum grid value).

peak for the 15% and 20% thresholds is similar, but the 10% threshold creates significantly larger areas. Therefore, we chose the 15% because it was a sweet spot between POD and size of the areas (i.e., reducing false alarm areas).

The algorithm then selects the PST with the highest probability of detection (POD) for the day, and then evaluates subsequent PST areas based on POD. The set of rules the algorithm uses to select subsequent PSTs includes i) There is no more than a 10% spatial overlap with other selected PSTs, ii) the POD that is added with the additional PST area is greater than 10% (this rule ensures that there is a sufficient number of reports in each PST), iii) the area of the PST is at least 38,400 squared km, and iv) there are no more than 4 boxes

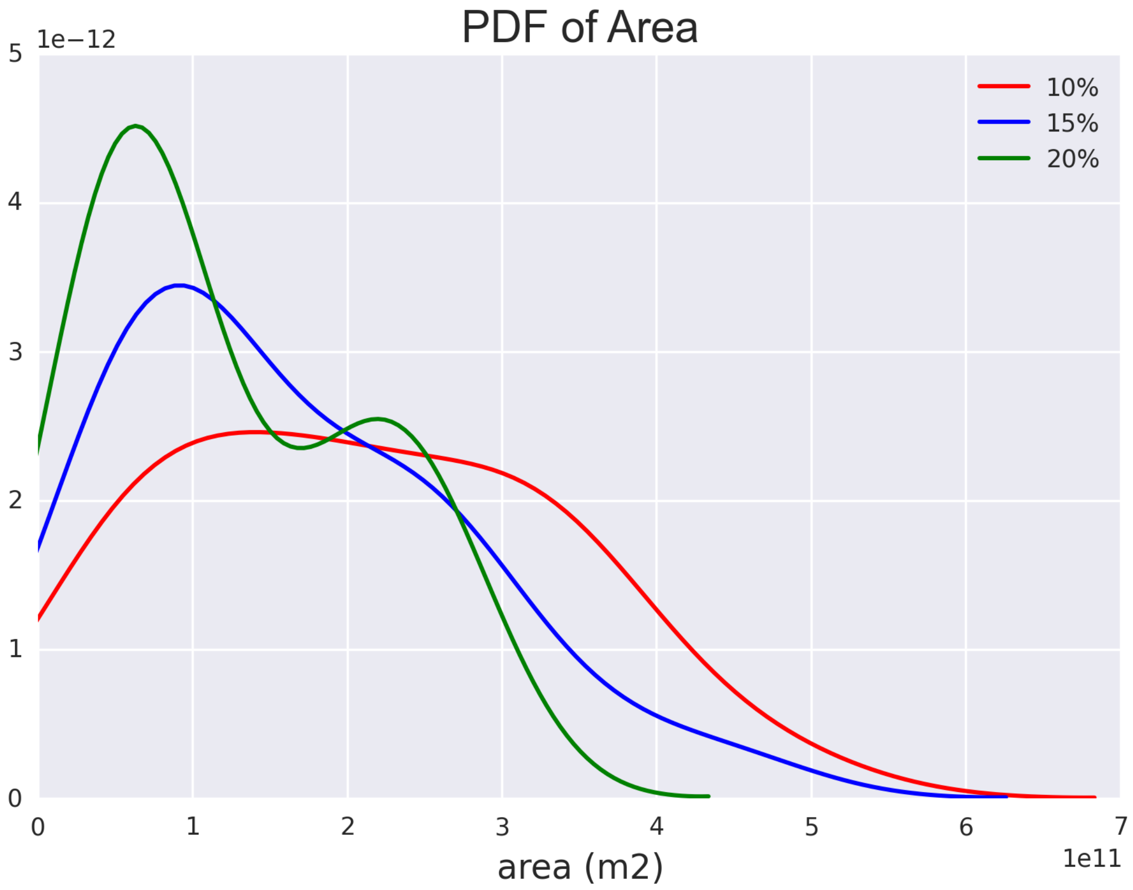


Figure 3.15: The probability density function of PST area for three different thresholds (noted as a percentage of the maximum grid value).

on a single day. These rules are applied to attempt to mimic what forecasters would draw if they knew exactly when and where reports were going to occur. Ultimately, the automatic PSTs should be the "forecast to beat".

During the evaluation of the previous days forecast, participants were asked to not only evaluate the quality of the forecast, but also comment (during discussion periods and survey feedback) about the benefits and challenges of the PST product. Survey instruments can be found in appendix A.1.1.

Finally, forecasters were also asked to fill out an exit survey at the end of their participation week. This survey asked questions about the PST concept broadly, including the perceived value and use by stakeholders. This is also where we asked forecasters about their opinions on overlapping PSTs, including interpretation and necessity. We asked very similar questions of the EMs to deduce if and how their interpretations differed from that of forecasters. Examples of questions asked on the forecaster end-of-week survey can be found in appendix A.1.2.

On the end user side, there were two experiments during 2019. The first occurred from 13 May to 24 May and included eight participants over the two week period. Participants ranged in jurisdiction from small counties to entire states, from utility companies to hospital networks (see appendix B.2 for participant jurisdictions). The second experiment included an entire integrated warning team, which means that the experiment included forecasters, EMs and broadcast meteorologists all working on similar simulated cases. This experiment spanned 3 weeks, starting 7 October and ending 1 November. There were a total of 11 EM participants over the three weeks, spanning jurisdictions from rural counties in Iowa to urban cities in New York (see appendix B.2 for participant jurisdictions).

Overall, there were some relatively large changes to the schedule and data collection methods implemented in the 2019 experiments. First and foremost, instead of running two simulated cases each day (see table 3.1), we chose to run one and include forecast information from day 4 down to the warning scale (see table 3.2). This choice was made to more adequately simulate the time periods when EMs are generally making decisions. Previous experiments revealed that EMs couldn't provide much feedback during warning operations because most of their decisions would have been made prior. This new schedule allowed researchers to more comprehensively evaluate the EM decision making process and challenges associated with the current product structure. Therefore, there were three cases tested in the spring experiment and three cases tested in the fall experiment. The spring cases included a supercell case in the Topeka, KS county warning area (1 May 2018), a messy tornado case in the Columbia, SC county warning area (24 May 2017), and a complex wind, hail, and isolated tornado case in the Goodland, KS county warning area (25 May 2017). The fall cases included a quasi-linear convective system tornado case in the Jackson, MS county warning area (18 April 2019), an isolated supercell tornado case in the Des Moines, IA county warning area (19 July, 2018), and an overnight severe weather and tornado case in the Wilmington, OH county warning area (27 May 2019).

The PST itself changed slightly for the 2019 end user experiments. First, the simulated PSTs were limited to the 15% area to reflect the forecasters' preference to only forecast for areas with a 15% minimum probabilistic forecast (Figure 3.16). Additionally, the colors of the timing areas were changed to a monochromatic scheme after feedback from the 2018 experiments indicated that the different colors can be interpreted as being different intensity levels (Figure 3.16).

In addition to including products with longer lead times, researchers also implemented some new survey strategies. The 2019 pre-test surveys remained similar to the 2018 experiment, which included questions about what SPC products participants regularly use and how their job changes when a weather watch is issued. The post-test was changed in 2019 to include some comprehensive questions about the PST concept and ease of use. The spring 2019 post-test included just a few questions, while the fall 2019 post-test included additional questions that separated different aspects of product ease of use. For example, the fall post-test included separate questions about whether or not the PST delivered

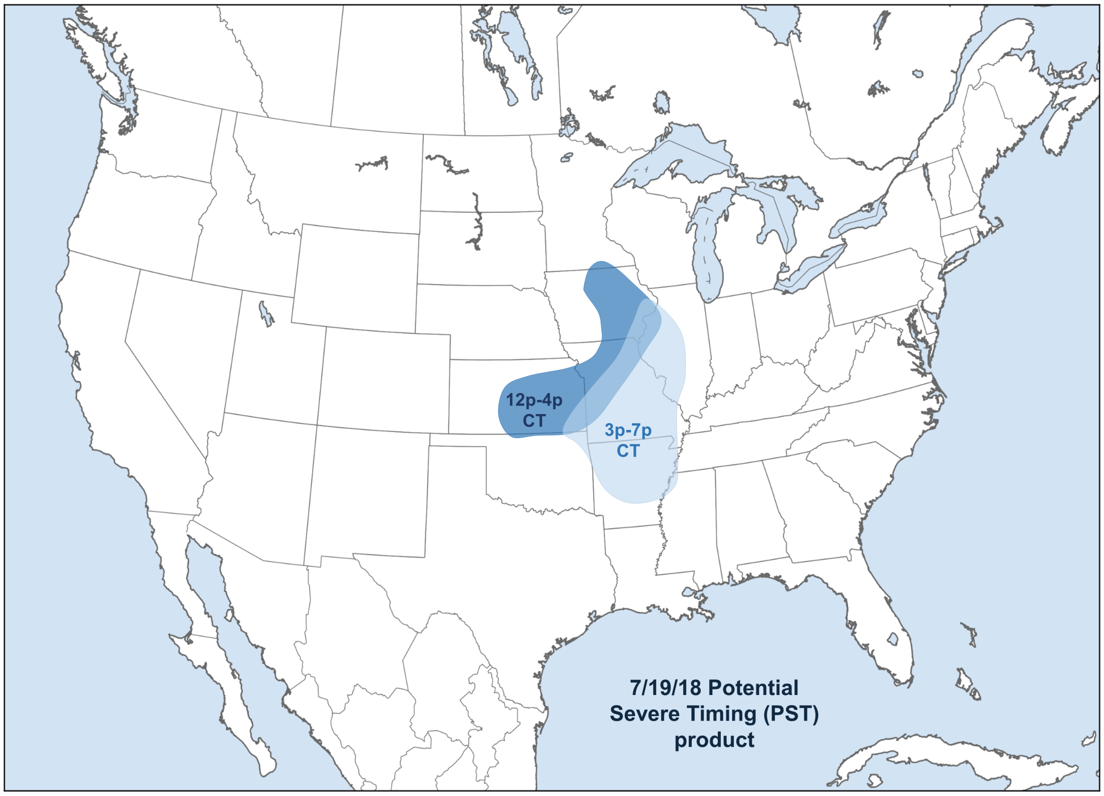


Figure 3.16: An example of a simulated PST for 19 July 2018 used in the 2019 end user experiment.

information quickly, was easy to use, and increased confidence in decision making (see appendix A.2.3).

The major difference in survey strategy implemented in the 2019 experiments were the *micro-surveys*. These instruments were very short and were administered 5-7 times during the simulated cases, generally after they saw a few operational products or a single prototype product (see appendix A.2.2 for an example of the PST micro-survey). The point of these micro-surveys was to ascertain the marginal benefit of individual products. Without regularly asking EMs what decisions they are making, researchers were previously unable to attribute changes in decisions or actions to individual products or new pieces of information.

Finally, one of the most important aspects of the end-user experiment is the open discussion period that follows each micro-survey. Researchers generally allotted 20-30 minutes of discussion time after each survey to provide context that generally would not be included in short survey answers. While we did have a set of questions to ask (see appendix A.2.4), we often allow the participants to drive the discussion when important topics were mentioned. We wanted this time to be lead by the participants, as it often led to questions, issues, or benefits that the researchers had not encountered previously.

3.6.2 Lessons Learned

The 2019 experiment year allowed researchers to continue to refine the PST product. After discovering some of the major challenges in 2018 (like overlapping time periods), we could focus on more nuanced issues, like forecaster training, guidance products, user visualization changes, and even best issuance times. These lessons are important to recognize *before* a product becomes operational so that conflicting strategies or interpretation methods can be streamlined, ensuring a consistent, recognizable, useful product is available on day one.

Forecasters in the 2019 SFE largely found the PST product to be a valuable use of time and thought that there would be many parties interested in the information. The 2019 SFE

PSTs performed well relative to operational products (like watches), with a POD of 0.65 and a success ratio of 0.27. A majority of the participants (32 out of 42) who answered the question agreed or strongly agreed with a statement discussing the value of the PST (Figure 3.17). In survey responses, forecasters recognized what researchers hoped they would see; that the PST product is an extremely simple product by design. It is meant to be a product that partners can look at for a few seconds and take away a useful piece of information.

"That was my first time to generate PST. I think the tool is easy to use."
"The key for me is the construction of a convective story line - what storm modes will occur where and when? The PST is then a visualization of that story line."
"[The PST] could potentially add an enormous amount of value to forecasts."

Ultimately, this is a reaffirmation of what researchers and operational meteorologists already knew; partners and members of the public often have questions about the timing of severe weather and use that information to make decisions. However, with simplicity comes challenges conveying all of the pertinent information, like the uncertainty of a forecast. This subject was brought up numerous times by both forecasters and emergency managers, and was a topic that researchers had not considered previously. Some forecasters even suggested a sister product under development in the SPC, called the temporally disaggregated probabilities. This product consists of the convective outlook probabilities at 4-hour intervals, which displays more information, but is not as intuitive or straightforward as the PST. Future iterations of timing products should consider how the strengths of both products could be manifested into a single product. Sample quotes from survey responses are quoted below:

"One big weakness is that the PSTs have difficulty displaying uncertainty."
"If [the PST] is given to an audience who doesn't understand weather forecasting, they could take it to mean that the threat ends right at the end of the period, when in reality, there is uncertainty."

”The [PST does] provide more timing information which is useful to EMs, public, etc., but it seems like those could be more easily provided by the temporally disaggregated products.”

Although the forecasters see the value of the PST product, they also recognize that the creation of the product is not always easy. In fact, a slight majority of the forecasters (75 of 143) reported that the previous day’s forecast was difficult to create (Figure 3.18). While this result may be concerning, we believe there is valid reason for the creation difficulty. First, it is important to remember that this process is not part of the current forecasting routine. In theory, this product would be easier to create if forecasters were doing it regularly. Second, most of the participants in the SFE are local forecasters or developers, meaning they do not regularly create products on a national level, which likely also contributes to their difficulty in creating the product. Finally, it would also be interesting to know what the baseline difficulty is for products that are regularly created. Does the PST rank higher than other products? Lower? Without this baseline knowledge, it is difficult to say whether or not the PST is unreasonably difficult to create.

Although the forecasters reported challenges with creating the PST, they also provided some useful suggestions to decrease those challenges. In survey responses, forecasters reminisced upon how the evolution of their forecasting strategies. Some used model forecasted updraft helicity tracks alone to make their PST areas, while others wanted a slew of model parameters and aggregate guidance products. These statements reinforce the knowledge gained in 2018 that there cannot be a single forecast process, as each individual builds their own process that varies drastically between forecasters. Some forecasters go so far as to want information about a model’s performance so they can appropriately assign weights to each model they consider. Some selected quotes from survey responses are noted below:

”I did wish I looked at more than just the UH tracks while compiling the PSTs”

”I would like to be able to verify the performance of different models before deciding on how I use them to produce the PSTs.”

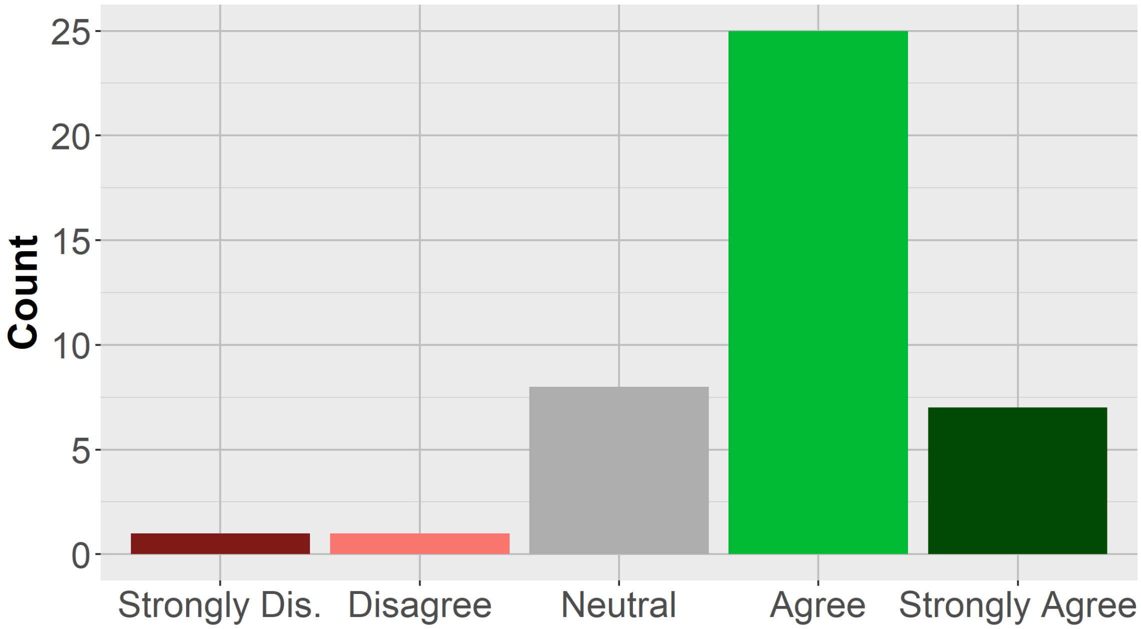


Figure 3.17: Forecaster response to the survey prompt: ”How much do you agree or disagree with the following statement: The added value of the PST product is greater than the added workload”.

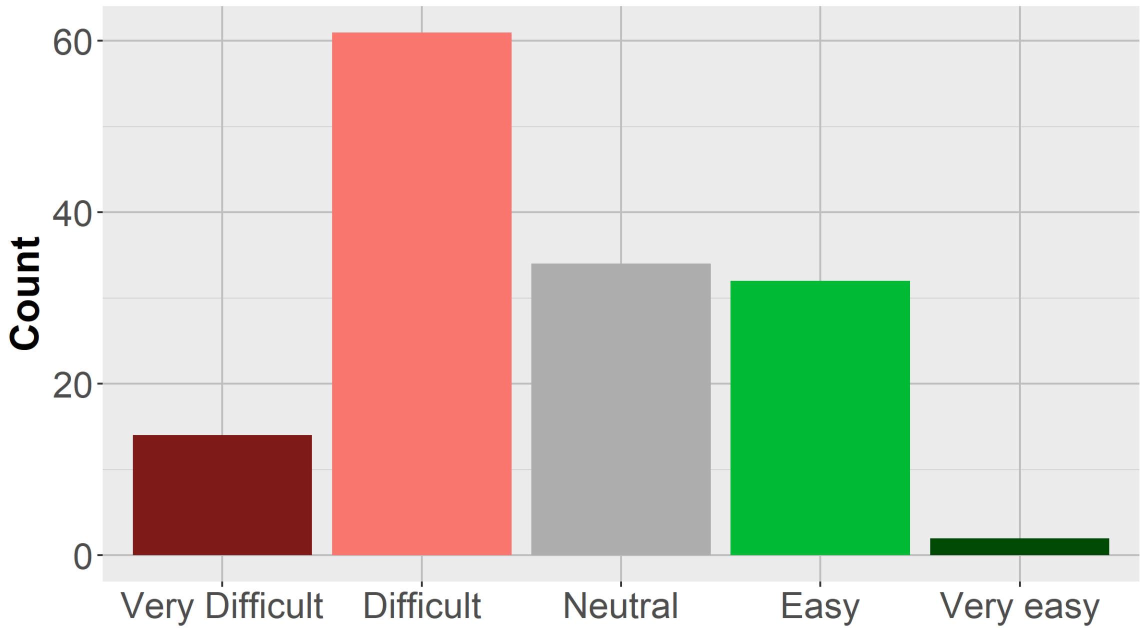


Figure 3.18: Forecaster response to the survey question: ”How difficult was it to create the PSTs yesterday?”

Forecasters also got to evaluate a new guidance tool in 2019; the *automatic PSTs*. While the feedback was generally positive for the concept of a first guess PST, the current iteration leaves some questions to ponder. Most forecasters agreed that having the automatic PSTs was at least useful to glance at (Figure 3.19). However, the ensemble subsets were difficult to work with because forecasters were unaware of the model biases or were uncomfortable relying on a single configuration. Therefore, future testing should allow forecasters to use whatever guidance they wish to. It would be even more useful if researchers could capture (or if forecasters could report) what guidance they were using.

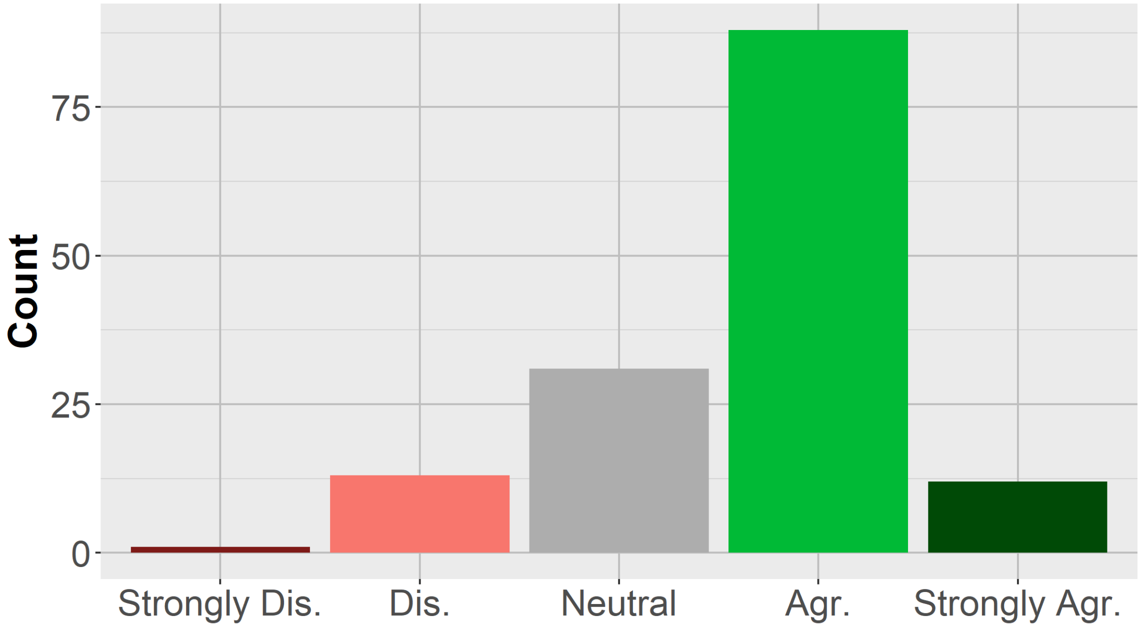


Figure 3.19: Forecaster response to the survey question: ”Was the ”first guess PST” guidance product from your ensemble subset useful?”

In addition to the first guess guidance, forecasters provided examples of other information they would ideally have access to while drawing a product like the PSTs. They include examples such as high reflectivity, updraft helicity thresholding, and even some experimental products like the Warn-on-Forecast (WoF) products (Skinner et al. 2018). These suggestions should be evaluated with expert forecasters and included in future experiments to evaluate their effectiveness. Selected quotes about additional guidance products are below:

> *"I'd like to have auto-drawn [PSTs] from some thresholds from all modelling systems shown separately."*
>
> *"I often think that it would be useful to have some guidance from model verification statistics, to understand for example the likelihood of the model missing the triggering."*
>
> *"I'm not sure. It seemed that I could have had too much information as opposed to not enough."*
>
> *"WoF guidance might have been helpful. The eastward speed of progression of the squall line across LA was not well captured in the guidance."*
>
> *"Time of UH first exceeding a threshold, time of last exceedence. These can get noisy with multiple waves of convection but are still useful."*

Discussion periods and survey feedback provided researchers with a wealth of feedback to consider in future experiments. A majority of this feedback was in the form of challenges that forecasters recommend remedying before the product becomes operational. These challenges can mostly be categorized into one of the following four categories: personal, guidance, meteorological, and experimental. It is useful for researchers to categorize these challenges to identify the most pressing issues from those that can be remedied with training or experience.

The personal challenges are mostly comprised of the latter, that is, challenges that would mostly be fixed with experience creating the product. Examples of personal challenges include: inexperience, lack of confidence in forecasting abilities, and internal calibration (i.e. knowing when a forecast is "as good as it gets"). One personal challenges that stuck out as needing more training or better tools was related to synthesizing the mountain of information into a single forecast. Some forecasters thought that this was especially challenging given the need to understand the spatial *and* temporal evolution of the storms.

Researchers may be able to construct a tool or guidance product that more effectively synthesizes this information, which should be explored in future iterations of the experiment.

Technical or experimental challenges may also be outside the realm of control of researchers or participants, but some of the suggestions should be considered in future experiments. One such suggestion is the amount of time given to forecasters to create the PST product. During the 2019 experiments, participants got anywhere from 15-45 minutes to draw individual PSTs. While the longer end of that range may be sufficient, the shorter end is almost always not enough time. There were many comments about the lack to time to synthesize the information in the PST, which should be strongly considered in future experiments. In addition to the time crunch, some other technical nuances made drawing the PSTs challenging. For example, a few forecasters mentioned wanting a larger group to draw the PSTs so they were not relying on their personal expertise alone. Since there are benefits to deriving individual forecasts, this suggestion should be considered along with the information that will be lost. Other forecasters wanted to have the ability to change their forecast as new information came in, or adjust the length of the time window for a single PST. Again, these suggestions would fundamentally change the PST concept, so they should be considered along with the information lost. Finally, a few participants mentioned not being able to decide between drawing larger PST areas that will capture more reports but have higher false alarm area, or drawing smaller PST areas that won't capture as many reports, but also won't have as much false alarm area. This is an example of an opportunity to have forecasters, users, and social scientists come together to discuss the benefits and costs of these two options. This exercise would likely be beneficial to all parties involved.

Challenges related to guidance products and meteorological uncertainty will likely always be present, but there are suggestions from the 2019 experiment that could limit these challenges in the future. Many participants noted the spread in model solutions caused major discomfort in choosing PST time frames. Ideally, as model skill continues to improve, the spread between model configurations will decrease. Without this increase in skill, a measure of model performance in the short term may help forecasters decide which models to hedge towards or away from. Other meteorological challenges like knowing the exact frontal placement in 12 hours may never be completely placated, but being aware of the current state of forecast skill and understanding that timing information is sometimes more challenging to produce than spatial placement is important for researchers to keep in mind when considering forecaster workload and fatigue.

While forecasters did note numerous challenges when creating PST information, they also generally recognized the value for of timing information for forecasters, the public, and especially partners (Figure 3.20). This information is important to verify to ensure there is motivation to create the product, particularly when it is a challenging task.

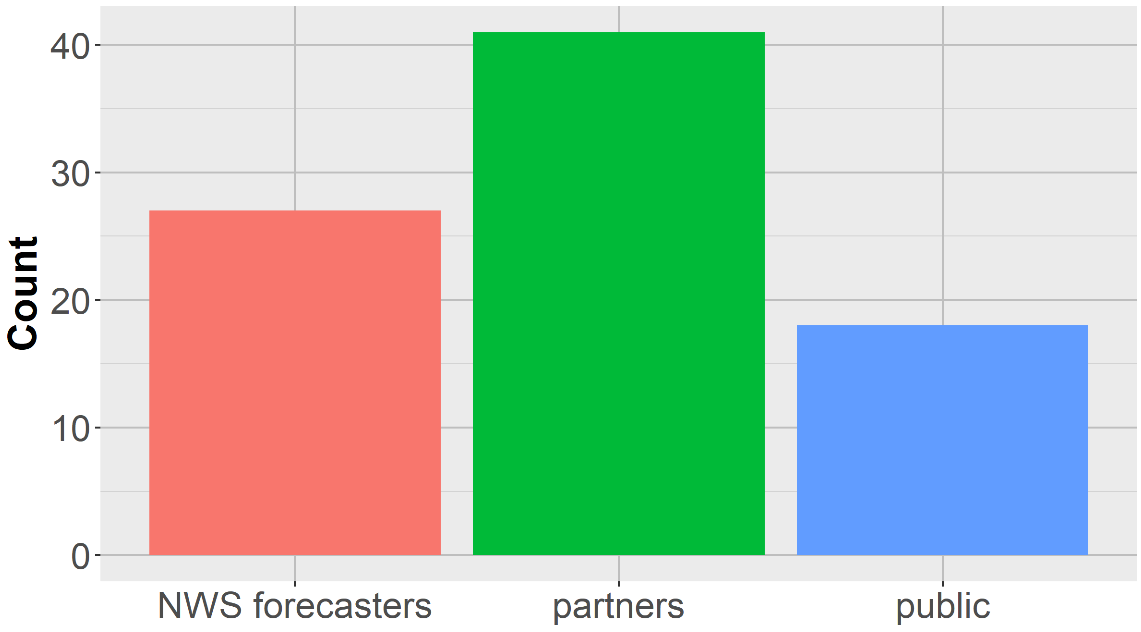


Figure 3.20: Forecaster response to the survey question: ”Who do you think the audience for the PST products should be?” Respondents were allowed to check more than one option.

During the end user experiment, EMs once again expressed the need for convective timing information. They thought the the PST was overall easy to use (Figure 3.21), increased

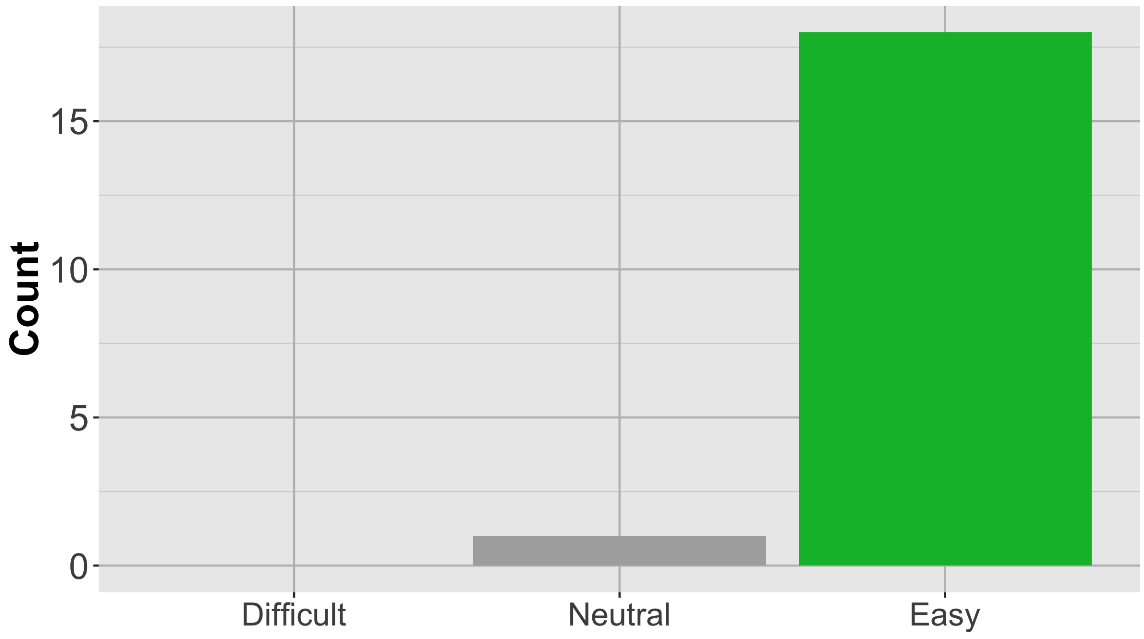


Figure 3.21: End user response to the survey question: Please evaluate the degree to which the PST: - Was easy to use”.

confidence in decision making (Figure 3.22), and would be used regularly (Figure 3.23). Of the multiple prototype products tested in the 2019 end user experiments, the PST was most often the product that EMs wanted to see operationalized first because it is simple to use and it very closely resembles products that are currently available. Many participants noted that the PST would fit in with their current workflow seamlessly because it compliments the convective outlook, which they are already looking at.

”The PST complemented the outlooks perfectly. It was a really good addition. It was a simple graphic and it was exactly what I thought I was reading” -EM 10

”Everyone wants to know about time.” -EM 23

”If you’re concerned with people, then you need to know the when.” -EM 27

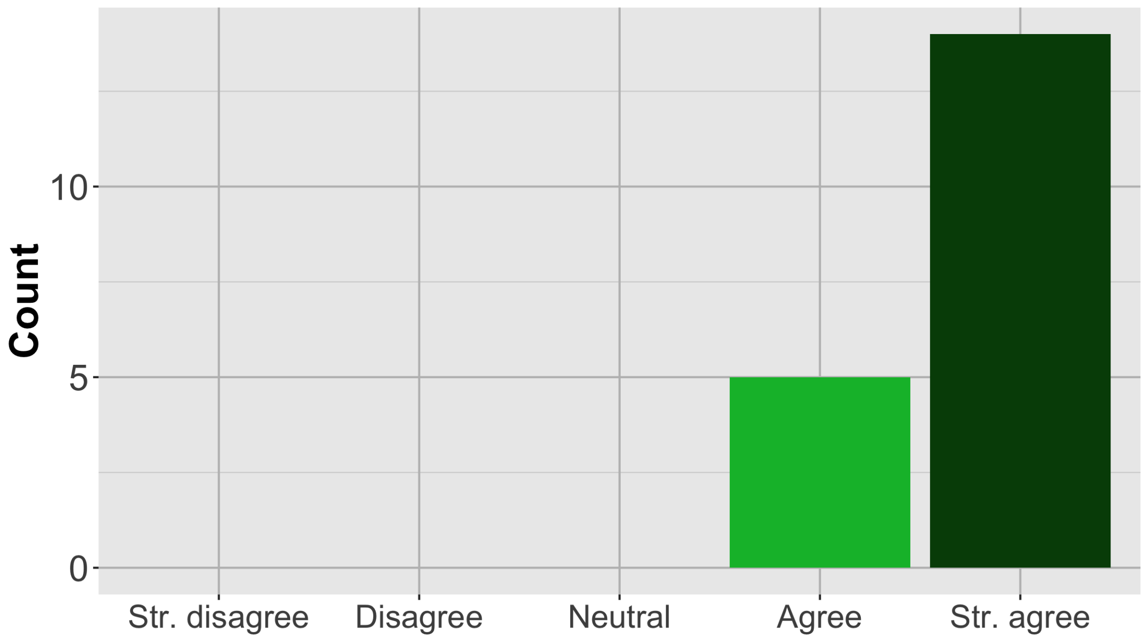


Figure 3.22: End user response to the survey question: ”Please evaluate the degree to which the PST: - Made you more confident in your decisions”.

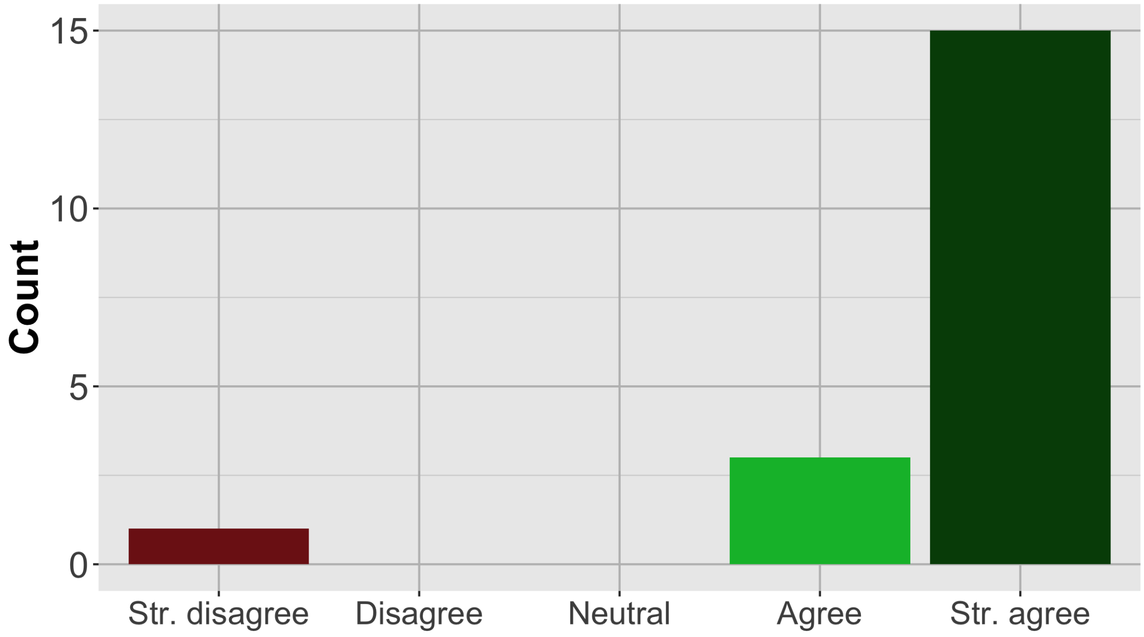


Figure 3.23: End user response to the survey question: ”How much do you agree or disagree with the following statement? I believe I would use the PST on most severe weather days for my area”.

Beyond just the ways to incorporate PST information into their own workflow, EMs could also imagine situations when the timing information would be useful to other stakeholders, like school officials or transportation workers. Participants also thought that this official timing information would provide the means *and* justification to allocate resources or change staffing schedules before the event begins.

”School districts don’t always listen to me. Every now and then they call me and want to know when they’ll be impacted. This would help provide the missing piece.” -EM 16

”[Schools] need to call in bus drivers, notify parents about kids getting released/held. A lot of storms come in during that time frame (2:30-4:30) so this would be beneficial to get earlier in the day.” -EM 22

”This would really affect my staffing. I can call my volunteers and give them the specific time I need them. It’s also helpful because ’afternoon’ means something different to everyone.” -EM 12

”Time frame is key for a number of reasons. If it’s earlier in the day I have a lighter response because of volunteer firefighters are at work. If it’s the end of the day, I will have more work.” -EM 11

”It helps with staffing at the state level.” -EM 14

Importantly, the 2019 end user experiments illuminated some less-known individual differences between EMs. For example, researchers assumed that participants would know about and have experience using SPC products. However, some participants noted that they almost never use SPC products in their daily jobs. Even more prevalent was the ”wait until something peaks my interest” mentality. Most participants agreed that they don’t monitor the SPC outlooks every day. They will keep an eye on the weather through other means (like local broadcasts), and when they note an upcoming event, *then* they will go look at SPC products. This was somewhat surprising to researchers and reinforces the idea that we need to be more aware of the differences in workflow to try to tailor the experiment more closely with each EM. The feedback from participants will cease to be authentic if we are asking them to learn a new workflow *and* evaluate prototype products.

Along similar lines, researchers spent time in the 2019 experiment evaluating details of the product structure with end users, including who they think should issue products and when they should be issued. Participants were mixed regarding whether the local NWS office or the SPC should issue the PST. There was some consensus that the SPC could issue the first PST and the local office (with knowledge of community structures) could refine the information. While this may be an ideal scenario, there may not be the working hours to have multiple institutions touching a product. Additionally, this dual-creation scenario may set forecasters up to disagree and have inconsistent messages. Other participants (particularly those from Oklahoma) thought that the jurisdiction should stay solely within the SPC.

"The NWS knows our local community, they can tweak what the SPC puts out and then give it to [EMs]." -EM 20

"The national office looks at the continent and they're making continental interpretations of probabilities. Local office doesn't do that. They rely on [SPC] products to make their tools. I don't think they should [create the PST]." -EM 21

Ultimately, one of the biggest takeaways from the 2019 end user experiments was the strong desire for simple, straightforward information. This is where the PST often stood out from other products. The PST shows one piece of information directly on the map (i.e., there is no legend that must be interpreted for timing information) in a clear way. Researchers like to think of this as the "3 seconds from 3 feet away" rule. If EMs cannot get the information they need in 3 seconds while standing 3 feet from the screen, they may not use the product. Given the immense number of tasks EMs must juggle, it is no surprise that difficult-to-interpret products are not utilized as often.

While there was plenty of positive feedback for the PSTs, participants also detailed some changes that would either improve the clarity or utility of the product. First and foremost, if there are going to be colors used (which helps with visualizing the different areas), then participants almost unanimously wanted information to explain what the colors meant. In the simulated PSTs (e.g. Figure 3.16), the colors simply denoted different timing areas. However, some participants interpreted the colors to mean confidence or even different hazard type. In the future, researchers should include a legend or written description of what the colors mean.

Other additions to the PST should be considered closely along with simplicity. While many users want additional information, they also want the product to remain simple and easy to interpret. Some ideas for other information to include in the PST are; hazard types, confidence in the time frame, and geographic markers. A few of the weather savvy EMs wanted to see hazards included in the PST product because they know that a lot of systems start as a hail and tornado risk and evolve into more of a wind risk over time. Displaying the hazard types with the timing information would help them prepare for the specific threat. Confidence in the time frame was more often requested than the hazard types. Almost all of the participants noted inferring confidence from written forecast discussions or through conversations with local forecasters. This confidence level is very important to them and often impacts the decisions they make. Many of the participants interpreted the colors on the PST to be confidence, with darker colors being more confidence and lighter colors being less. Admittedly, this thinking did play a part in the chosen color scheme, as earlier times (i.e. closer to when the PST was being created, and therefore carrying potentially more confidence) were colored darker and later times were colored lighter. It is not unreasonable to see how participants also came to this conclusion, but this tactic must be formalized in future iterations of the product. Finally, EMs use geographic markers (like county lines, highways, cities, etc.) to make inferences about the risk to their jurisdiction and thus decisions about how to prepare or respond. Including the option to toggle these markers on or off (like most SPC products) would be helpful for EMs to make more authentic decisions during the experiment.

”I’d like to see hazards included in the times. It helps because events change over time. Could go from supercell to a wind event.” -EM 19

”I’m a confidence level guy. Or a percentage guy. If that could be added in it would be good. I don’t know if you could add that to the colors without making it too busy, but confidence in the timing itself would be nice.” -EM 20

”I’d like the ability to toggle county outlines.” -EM 18

Another open question remains about what time would be best to issue a product like the PST. While many participants want the product to be issued earlier, they see a trade-off between earlier issuance and confidence/accuracy. Would an earlier product be significantly less accurate? Would the forecasters be less confident in their timing information? These are just some of the issues that EMs consider when using weather products to make decisions. More work will need to be done to evaluate the effectiveness of a product that is issued earlier in the day.

Finally, researchers revisited the question about overlapping PST areas with both forecasters and EMs. Not surprisingly, there is no consensus on what to do with overlapping areas, but everyone seems to have an opinion. Many forecasters thought that no overlap should be a goal to strive for, but that there are many days that almost *require* overlap because of the large amounts of uncertainty. One item that forecasters agreed with was the need to understand how users interpret overlap before deciding a forecasting best practice.

”I specifically avoided overlapping areas both spatially and temporally. This made the forecast worse in comparison to the best forecast.” -Forecaster participant

”It appeared that there would be multiple rounds of severe weather lasting longer than 4 hours, so it was necessary to overlap PSTs.” -Forecaster participant

”The overlap in the [PST] areas is still a point to ponder. Specifically, we still haven’t answered how emergency managers would interpret such a product.” -Forecaster participant

Also not surprisingly, EMs had mixed opinions about overlapping areas. Some thought that the overlapping areas made logical sense because they showed movement in the system. Others, especially when prompted to decide the time frame for point locations in the overlap area, could not reconcile what the overlap actually represented. Regardless of consensus, the fact that there is different ways of interpreting overlap indicates that it is not good practice to include it. The more room for individual interpretation, the more confusion will likely ensue. Future research should investigate the decrease in accuracy (if there is one) when overlap is removed to decide if/how overlapping areas should be included.

"I inferred some movement. I thought the other bubble meant it ended at 6pm" -EM 16

"It made sense." -EMs 9-12

"I'm confused by the overlap, I'm looking at timing and location" -EM 20

"The overlap was a representation of a slight movement in time. I interpreted it as the middle of both portions." -EM 15

"I'm not sure how to interpret that...is that where...I don't know." -EM 23

"I get what [the forecaster is] trying to convey-as something moves, how do you keep the window moving with it? I think the overlap [area] is too small to have a 3rd category. So this is the next big thing. If it was large, then this would be okay." -EM 19

"I kind of like that there's overlap. I don't expect anyone to have a crystal ball so I don't expect a clear delineation line. But what that does is help me answer questions from the school district. When you factor in the overlap, I'm estimating our area looks like 3-7pm. I'm putting an hour left and right. It's better than 'sometime today.'" -EM 21

In an attempt to clarify the interpretations of overlapping areas, both forecasters and EMs were asked to identify the timeframe they thought was forecasted for the red star in Figure 3.24. Perhaps not surprisingly, results showed nearly even representation of each of the four response options *and* many different write-in options. While this did not clarify what should be the overlap policy, it did further highlight the multiple interpretations and the likely need to greatly reduce or eliminate of the use of overlapping areas.

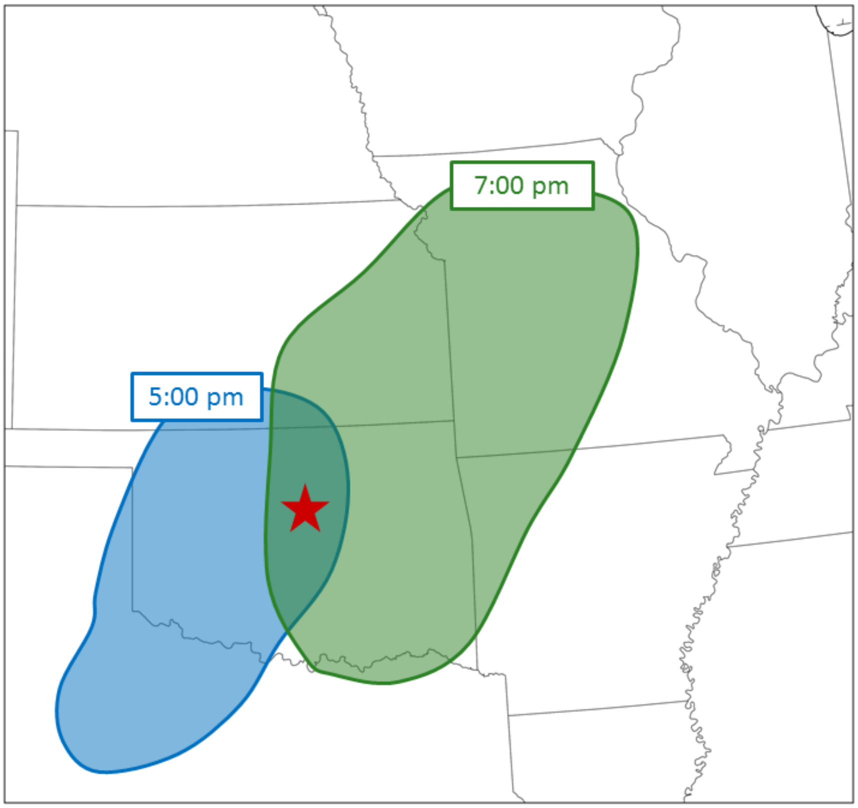


Figure 3.24: The figure shown during post-test surveys discussing overlapping PST areas. The question wording was: ”Below is an example of a forecasted PST area. Please indicate which timeframe is forecasted for the red star.”

Ultimately, the 2019 experiment highlighted again the value of timing information and the need for a product to be simple and easy to interpret. The limiting of the PST areas to the 15% area makes practical sense for forecasters and aligns with EMs interpretation of forecasting procedure. Future work should focus on clarifying color use, potentially adding in hazard or confidence information, and consider issuance times and who will create the timing forecast. Table 3.5 summarizes the biggest takeaways and future actions that should be taken.

Table 3.5: Lessons learned in the 2019 experiments and future actions to consider.

Lesson	**Future steps**
Automatic PSTs are useful but need refinement	Investigate higher UH thresholds to reduce the number of proxy reports. Consider other smoothing techniques.
Forecaster guidance is still lacking timing information	Consider adding in forecaster suggestions to the forecast drawing tool.
Time to forecast PSTs was still too short	Consider rearranging the HWT schedule to include *at least* 30-45 minutes of dedicated PST time.
EMs get much of their weather info from local offices	Present the PST as coming from the local office. Evaluate the trust in the forecast compared to trust with SPC issuance.
EMs use hazard information and confidence levels to make decisions	Consider having forecasters indicate confidence levels and forecasted hazards within each PST.
Color has many interpretations	Include a legend with the PST to indicate *what* each color represents.

3.7 Discussion

Ultimately, the years of testing longer lead time timing products with forecasters and EMs illuminated numerous benefits and challenges that researchers had not considered previously. What researchers believe will be an easy to create, easy to use product is often deemed too complex or trivial when the product is dumped into an operational setting. This work highlights the importance of rigorous review with creators and users of a product. What one group sees as beneficial may not be remotely useful or intuitive to another group.

The PST product is a very simple product that displays one piece of information: severe weather timing. While it may be easy to assume that this product was developed over the course of a few weeks, the iterations it took to get to the 2019 version are numerous. What started as a graphic with lines of time has ended up as an areal coverage product that shows the evolution of the system with enough flexibility to impart uncertainty in the placement of storms at future times (Figure 3.25). This product is reasonably straightforward for forecasters to create, and it meets a variety of needs for emergency managers. Namely, the PST has been shown to; aid in EMs decision making process, serve as a communication tool for them to report to management and end users quickly, and it is easy to interpret.

The co-production process that was attempted during the testing phase of the PST product highlights the benefits and challenges of co-production. In hindsight, it would have been beneficial to have forecasters and users in the room from the very beginning. Maybe then researchers could have skipped the isochrone visualization all together. Regardless, the process of involving all parties that will touch a new product should be the standard of practice in applied research settings. This will become even more important as NOAA and the National Weather Service move towards offering more decision support services (DSS). As forecasters are required to offer more and more tailored information for different user groups, it is imperative that those user groups are involved in the decision making process about *what* information is provided and *how* that information is displayed. The last thing the weather enterprise needs is to have scientists spending time creating information that ultimately no one understands or uses. An iterative, co-production process at minimum would illuminate issues with products before they become operational, and at best would generate well understood, anticipated products that are useful from day one of operational creation.

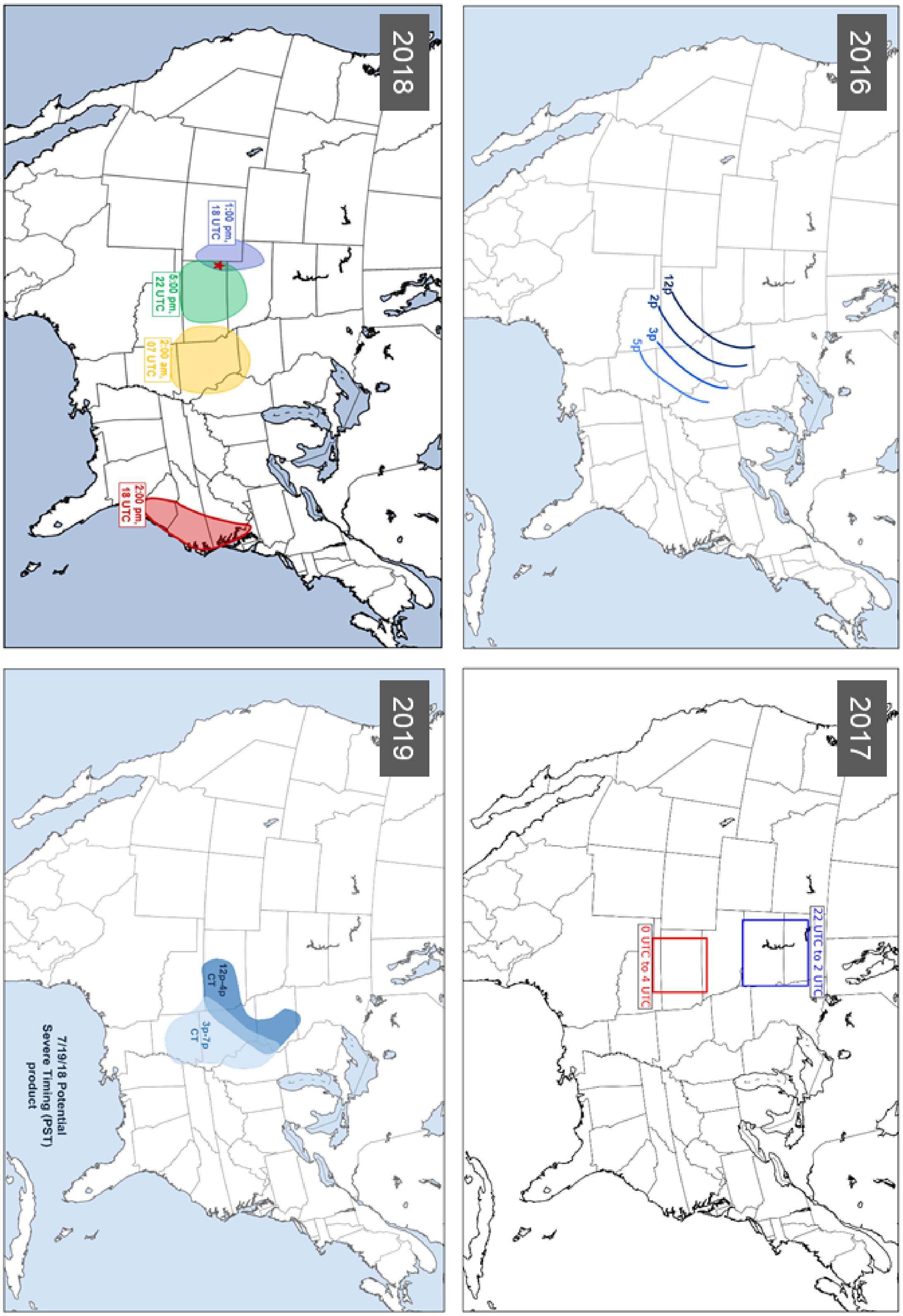


Figure 3.25: A summary of the timing product evolution throughout the testing period from Spring 2016 through winter 2019.

Chapter 3 items of note:

- Forecasters understand the value of timing information, but it can difficult to produce accurately and reliably.
- Emergency managers have a need for timing information hours (or even days) before the event occurs.
- Products need to be consistently produced so users can expect them and evaluate the accuracy and usefulness.
- The co-production of scientific information needs to be iterative and allow for significant refinements to the original concept.
- This product is the culmination of years of analysis, resulting in the most useful information being displayed in the most effective manner for users to find value in it.
- In addition to the creation of useful scientific products, the co-production process also develops a supportive user community and increases trust between forecasters and users.

Chapter 4

The impact of hours of advance notice on protective action in response to tornadoes

4.1 Introduction and background

As forecast technology continues to improve, we may start to see more specific forecast information (including timing information) earlier in the event timeline. This may mean that people could know their specific threat time frame 4 - 8 hours ahead of the actual event. A change like this would open up a new realm of potential response actions, many of which have not been studied. This work begins the process of understanding what types of response actions individuals may take given hours of notice to tornadic events, and how those actions differ from those currently taken given minutes of lead time for tornado warnings.

The NWS is the government entity tasked with issuing hazardous weather forecasts for the United States for the protection of life and property and enhancement of the national economy (NOAA National Weather Service 2019). Their suite includes products covering all hazard types, from air quality alerts to winter weather products. Specific to severe weather events, there are generally three different levels of products that comprise the public communication process. The first level includes convective outlooks, which are forecasted by the Storm Prediction Center (SPC) up to eight days in advance, and are generally on a regional or multi-state scale. The second level includes mesoscale discussions and severe thunderstorm/tornado watches, which are also issued by the SPC. These products are issued on the day of the event, generally 1-3 hours before the event begins, and are usually on a multi-county or statewide scale. Finally, the third level includes warnings, which are issued by the local NWS office. They are generally valid from just prior to the event occurring and are usually the size of a county or two.

While the current system includes three distinct levels of products, a proposed system from the Forecasting a Continuum of Environmental Threats (FACETs) project aims to provide a continuous flow of hazardous weather information (Rothfusz et al. 2018). Conceptually, this system would provide each individual user with information specific to their situation and threat tolerance. For example, this future system may supplement current products with a continuous stream of probabilistic hazard information (see Ling et al. 2015) that users can view at any point in time or space based on their pre-chosen alert-level settings. While potentially beneficial, some key partners (like emergency managers) often rely on specific products to make decisions or activate procedures (Cross et al. 2019). Likewise, tornado watches seem to improve the tornado warning process in local forecast offices (Hales Jr 1989). From a public perspective, watches often serve as the first line of defense to initiate protective action. Generally, the more severe the watch type, the more likely people are to stop their activities and start monitoring the situation (Gutter et al. 2018). These findings raise an important question: should certain products (or product levels) be maintained in the proposed FACETs system? As forecast technology continues to improve, the current products may start to evolve and serve a different purpose, but their existence may still be important to core partners and the public.

If some or all of the current product structure remains in place, the challenge for the FACETs paradigm then becomes developing a continuous flow of information while maintaining the current product structure of discrete forecasts. Currently, a multi-hour information gap may exist between the convective outlook and the first (if any) mesoscale discussion or watch, depending on the day. Although many NWS forecast offices provide information between these two products (often in the form of online or phone briefings and social media posts), there is currently no formalized product information available between the convective outlook and a mesoscale discussion or watch. One of the proposed solutions to help remedy this information gap is to include the time frame of severe weather along with the probabilistic and categorical risk levels in the convective outlook. An analysis of historical severe weather reports shows that a majority (greater than 95%) of daily reports occurring at a single location will occur within a 4-hour period (Krocak and Brooks 2020). Currently, the SPC forecasts the probability of severe reports occurring within 25 miles of a location over a 24-hour period. Since the analysis above shows that a majority of reports within 25 miles of a location will occur in a smaller timeframe (4 hours), the SPC could, in theory, provide information about that smaller timeframe (like when it will occur) without changing the definition of their probability forecasts. For example, the SPC could provide the categorical risk (marginal, slight, enhanced, moderate, or high risk) and a 4-hour time frame (1-5 pm, 4-8 pm, etc.) for each location.

While there has been little work conducted to understand response actions to hours of advance notice before an event, there has been some work related to warning-scale response actions. Most studies that ask participants about response actions consider immediate sheltering to be the most correct response (e.g. Jauernic and Van Den Broeke 2016). When evaluating the factors that change response actions, studies find that demographic characteristics such as education, age, and gender can impact sheltering behaviors. Responsiveness increases until about 65 years, then decreases with age (Chaney et al. 2013), increases with education (Balluz et al. 2000), and women generally seek shelter more often than men (Ripberger et al. 2015).

Other factors that impact response behaviors include the wording of the actual product. Impact-based warnings include language that discusses the potential consequences of the event, including damages and loss of life. Studies find that this type of language increases response actions, including plans to shelter (Casteel 2016, Casteel 2018, Ripberger et al. 2015). However, studies also find that even when residents plan to shelter, that often is not the first action they take because they will often confirm warning information from multiple sources (Jauernic and Van Den Broeke 2016).

Finally, many studies attempt to measure what fraction of participants respond to tornado warnings. Some include interviews after actual tornado events and find that anywhere from 43% to 79% of residents take action, depending on the region where the event occurred (Balluz et al. 2000, Miran et al. 2018b, Chaney et al. 2013). Studies that measure hypothetical situations find higher response rates, with anywhere from 75% to 90% of respondents claiming they will take action during a future event (Schultz et al. 2010, Lindell et al. 2016, Ripberger et al. 2015).

There is a fundamental difference between response actions for minutes of advance notice and those for hours of advance notice. Some research suggests there is a threshold of "too much" lead time on warning scales, or a point at which the threat no longer seems imminent and residents don't immediately head to shelter (e.g. Hoekstra et al. 2011, Ewald and Guyer 2002). In fact, one particular study finds that lead times of over 15 minutes may even increase the number of fatalities compared to an unwarned tornado (Simmons and Sutter 2008). However, while sheltering may be one of the only reasonable actions given 15 - 30 minutes of notice, there is a myriad of other actions that become more reasonable given hours of notice (i.e. leaving the area, monitoring the situation, preparing their home and family) that would in theory set other protective actions in motion (like preparing the shelter or important documents). This study aims to identify the actions individuals believe they will take given hours of advance notice for a tornadic event, and if (and how) those actions change given either four or eight hours of notice.

4.2 Data and methods

4.2.1 Survey data

The University of Oklahoma Center for Risk and Crisis Management (CRCM) fields a national survey to analyze public reception, comprehension, and response to severe weather forecast products (Silva 2017, Silva 2018). This survey, called the Severe Weather and Society Survey, has been fielded in 2017, 2018, and 2019. It utilizes an online format with a sample of US adults (age 18+) provided by Qualtrics, which maintains a large pool of participants that agree to take internet surveys. There were 3000 respondents in 2018 (the survey used in this study). Respondents generally took around 25 minutes to complete the survey and were compensated for their time. Dynamic sampling was employed, meaning participants were asked demographic questions before taking the survey, and were not asked to complete the survey if their demographic profile was already well-represented by the current pool of respondents. This process was used to ensure that the sample population was as representative of the US population as it could be. After the survey was fielded, responses were also weighted according to US Census estimates, further ensuring the results were demographically representative of the population.

One of the many unique aspects of the Severe Weather and Society Survey (hereafter WX18 for the 2018 iteration), is that there are two different types of questions employed. Some questions are recurring, where researchers attempt to establish a baseline of severe weather knowledge and response actions. Other questions rotate in and out, depending on what experiments researchers are interested in conducting each year. Although there were nearly 100 questions total on WX18, this study uses data from just a few different rotating questions to establish how more advance notice for the event (order of hours, not minutes) impacts tornado preparation and response actions.

The specific questions used in this study were open ended, meaning respondents could enter whatever information they like, and the responsibility to interpret their responses was placed on the researchers. Respondents were asked to describe what they would do given the knowledge that a large and dangerous tornado would impact their location in either four or eight hours. The amount of advance notice was assigned randomly to each participant, resulting in 1500 responses to four hours of advance notice and 1500 responses to eight hours of advance notice. Time of day was held constant at 9:00 AM to ensure that all respondents were anchoring to the same time of day and the activities that correspond with that time of day. After removing unusable responses (e.g. blank responses, random letters/characters), we were left with 1392 responses in the 4-hour category and 1404 responses in the 8-hour category for analysis. Differences in responses were compared to identify how the shift from four hours of notice to eight hours of notice would impact response actions. The exact wording of the survey question is noted below.

Imagine that it is 9:00 AM tomorrow morning and you are somewhat confident that a large tornado will hit your location in the next [RANDOMIZE: 4 or 8 hours]. What would you do? Please be as specific as possible.

Consistent with previous studies (e.g. Schultz et al. 2010, Ripberger et al. 2015), we use intended response actions as a proxy for actual response actions in this analysis. While there is little work that analyzes the relationship between intended and actual response actions to tornado warnings, there has been extensive work relating intended and actual behavior in other fields (e.g. Armitage and Conner 2001). This work shows that there is a significant link between intended and actual behavior, even when an individual is in a high stress situation (Kang et al. 2007). While it may not be a perfect proxy, we believe that our results provide some insight into what response actions might be given hours of advance notice for a possible tornado.

4.2.2 Response treatment

After fielding the survey, responses were divided into the two time categories (4-hours and 8-hours) for further analysis. We begin by comparing key word usage across the time categories. We do this by identifying the most common words that participants used and then compare the percentage of respondents that used each word across the time categories. For example, the percentage of responses that contain the word "shelter" in the 4-hour category is calculated as:

$$p_{shelter|4-hours} = \frac{n_{shelter|4-hours}}{n_{4-hours}} \cdot 100 \tag{4.1}$$

where n_shelter,4-hours is the number of responses in the 4-hour category that contain the word "shelter" and n_4-hours is the total number of respondents that were given 4-hours of notice in their response prompt. After those percentages are calculated, the difference in word use between 4-hours and 8-hours is calculated by subtracting the percentage of word use in the 4-hour category from the percentage of word use in the 8-hour category:

$$d_{shelter} = p_{shelter|8-hours} - p_{shelter|4-hours} \tag{4.2}$$

While the analysis of single words is a good starting place to understand basic response characteristics, the context of those words also plays an important role in understanding the actions people will take. To further investigate these response actions, all usable survey text responses are categorized into one or more categories. These 6 categories (shown in table 4.1) were chosen after reading the responses to ensure they encompass nearly all of the described actions. The categories are not mutually exclusive; in fact, many of the responses fit into multiple categories. Once categorized, the percentages for each category are calculated for both 4 and 8 hours of advance notice. Similar to the word analysis, the difference in these percentages is calculated to understand how changing from 4 to 8 hours would change respondents' actions. Finally, a two-tailed difference in proportions z-test is performed for all of the difference calculations to identify which, if any, of the word or category differences are statistically significant.

Table 4.1: Response categories and their descriptions.

Category	Description
Monitor	Watch as the situation unfolds, monitor apps or other weather information.
Prepare	Prepare for the incoming weather, gather family, supplies, etc.
Take shelter	Move to a safe area near the location the respondent is currently at and wait until the event is over. This includes locations that aren't specifically shelters (like basements or interior rooms), as well as specific storm cellars and shelters.
Leave	Leave the location the respondent is currently located. This includes responses that describe intention to leave the area to avoid the event altogether or to get to a shelter location.
Nothing	Do nothing in response to the event information.
Unsure	The respondent does not know what they would or should do in response to the event information.

4.3 Results and Discussion

4.3.1 Word analysis

The top words used in text responses are very similar for 4 and 8 hours of advance notice for a possible tornado (Fig. 4.1). In fact, 22 of the 25 top words used are found in both the 4-hour and 8-hour categories. The only difference is in the order of the most used words by number of times used. The three most popular words used with 4 hours of advance notice are “shelter”, “go”, and “monitor”, in that order. These words are found in 20.0%, 17.8%, and 16.8% of responses, respectively. Those same three words are also the most used with

8 hours of advance notice, except the order is "monitor", "shelter", and then "go". This time, they are found in 17.5%, 17.3%, and 15.9% of responses.

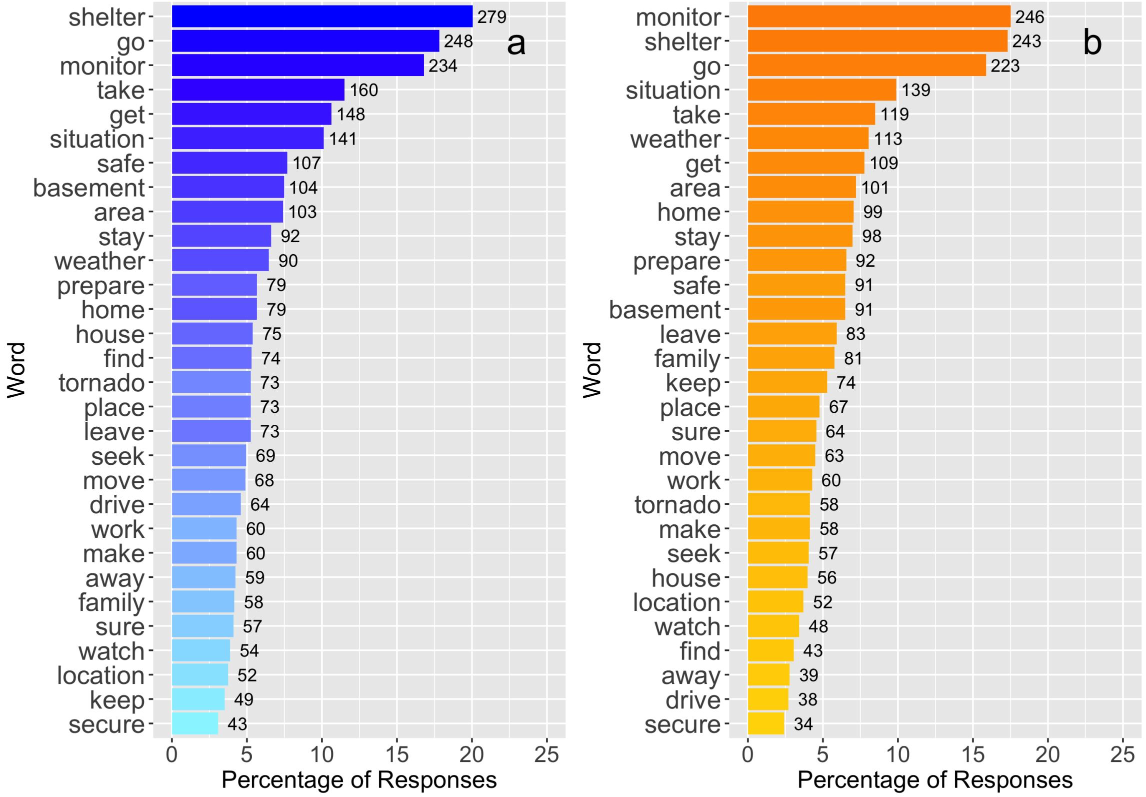


Figure 4.1: The percentage and number of responses containing the most common words in response to 4 hours of advance notice (a, n=1392) and 8 hours of advance notice (b, n=1404) for a dangerous tornado.

The differences between the most commonly used words suggests a pattern of more sheltering preferences with 4 hours of notice and more monitoring preferences with 8 hours of notice (Fig. 4.2). Words that relate to sheltering behaviors (like "shelter", "go", "take", "get", "safe", "basement", "find", etc.) are more prevalent in the 4-hour category, while words that relate to monitoring the situation or information gathering (like "monitor", "stay", "weather", "prepare", and "keep") are more popular in the 8-hour category. Due to the relatively low number of responses that contain the most common words ("shelter" appears in only 279 responses in the 4-hour category and "monitor" appears in only

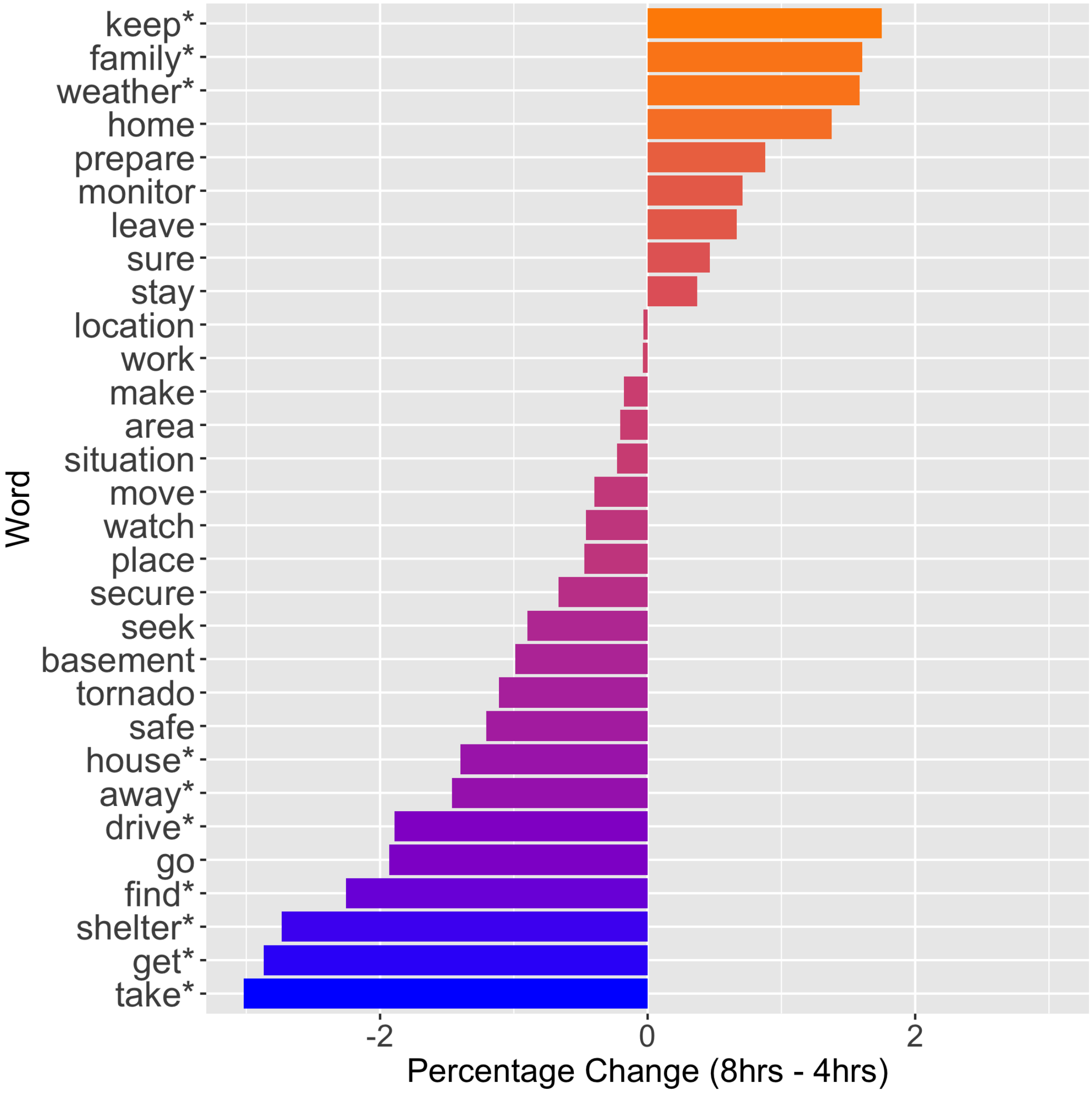


Figure 4.2: Difference in percentage of use between 8 hours (orange bars indicate higher percentage use) and 4 hours (blue bars indicate higher percentage use) of advance notice. * p < 0.1.

246 responses in the 8-hour category), many of the differences are not statistically significant. However, the words “take”, “get”, “shelter”, “find”, “drive”, “way”, and “house” are all more prevalent in 4-hour responses and statistically different from the 8-hour responses

at the p<0.1 level. This may suggest that there is a slight preference for sheltering behaviors within the 4-hour category. On the other hand, the words “keep”, “family”, and “weather” are more prevalent in the 8-hour responses and statistically different from the 4-hour responses at the p<0.1 level, which may indicate a slight preference for monitoring and preparatory behaviors in the 8-hour category.

Another common theme in both categories is respondents using the current system as an anchor for what they would do in a situation they have not been in before. Many responses displayed confusion or disbelief that there even could be 8 hours of advance notice for a tornado. Some responses reflect this disbelief, and then proceed to talk about what they would do with less time. This may highlight the need for education if a new system were to be put in place. If people are given some idea of what could be done with many hours of advance notice, they may be more likely to take precautionary actions.

While the percentages of individual words used are interesting and a good starting point, those words exist within context of the individual respondents’ situation. That context is often important when understanding what specific actions they plan to take. For example, the word “shelter” may be used more often in the 4-hour category, but it’s important to understand how it’s being used in both categories. With regards to 4-hours of advance notice, the word “shelter” is used mostly in a traditional sense; in statements like “I would go to shelter”. In the 8-hour category, it is often used to say that they would not head to shelter until necessary; in statements like “I’d monitor the weather and head to shelter when necessary”.

4.3.2 Categorical Analysis

Given that the context of the response (and not just the most commonly used words) is important for understanding common behaviors, we place each response into categories based on the most common response actions. Some of the same themes seen in the most commonly used words are also represented in the categorical analysis, but the categories also reveal actions that single words cannot (Fig. 4.3). As an example, prepare and monitor are the two most common categories in the 4-hour group (32% and 29% of responses, respectively), followed by sheltering (26% of responses). This is likely seen because the descriptions of preparing and monitoring do not necessarily need to include the words "prepare" or "monitor". For example, many respondents said that they would gather important items (prepare), head to a safe place (shelter), and then watch for updates on TV or their phone (monitor). None of those actions would have been captured in a word analysis, but become evident when comparing categories.

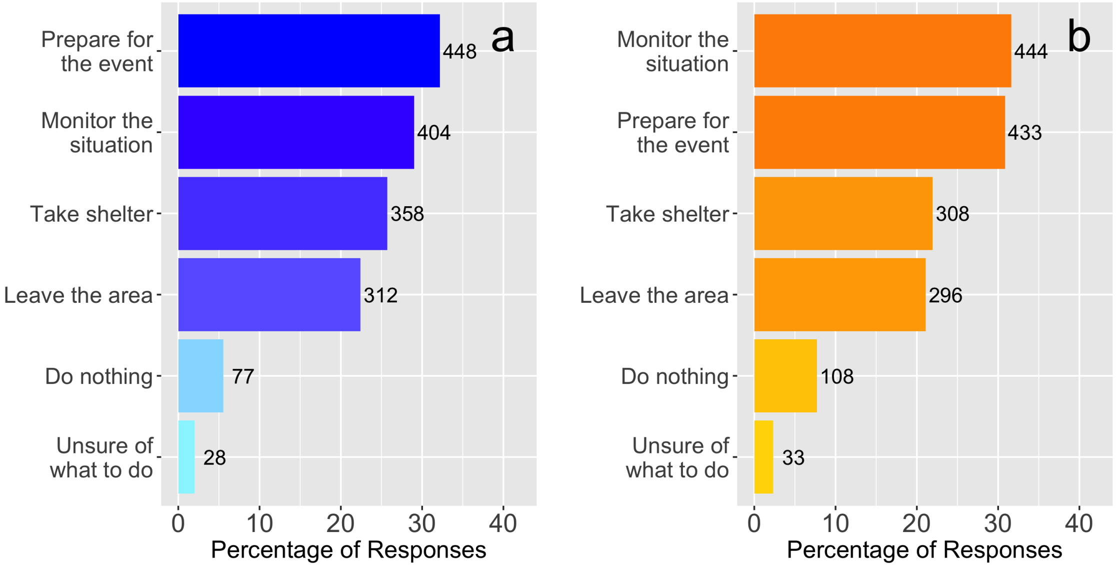


Figure 4.3: The percentage and number of responses in each response category for 4 hours (a, n=1392) and 8 hours (b, n=1404) of advance notice.

Respondents in the 4-hour group often mention gathering the most important documents/items and then sheltering with family. It is somewhat unexpected to see so many responses indicating they would immediately go to shelter when they would likely be in shelter for hours before the event occurred. In theory, they could get other tasks done or even leave the area before heading to shelter, but many responded as if they had mere minutes instead of hours.

The 8-hour group has similar response category percentages to the 4-hour group with just a few adjustments. The monitor and prepare categories switched places, with monitor being the highest category in the 8-hour group (32% and 31% of responses, respectively, Fig. 4.3). Many responses indicate that they would look for more information and act when the event was closer to occurring. The "nothing" category also increased to nearly 8% of responses, which may indicate that 8 hours was too much advance notice to begin taking precautions (Fig. 4.3).

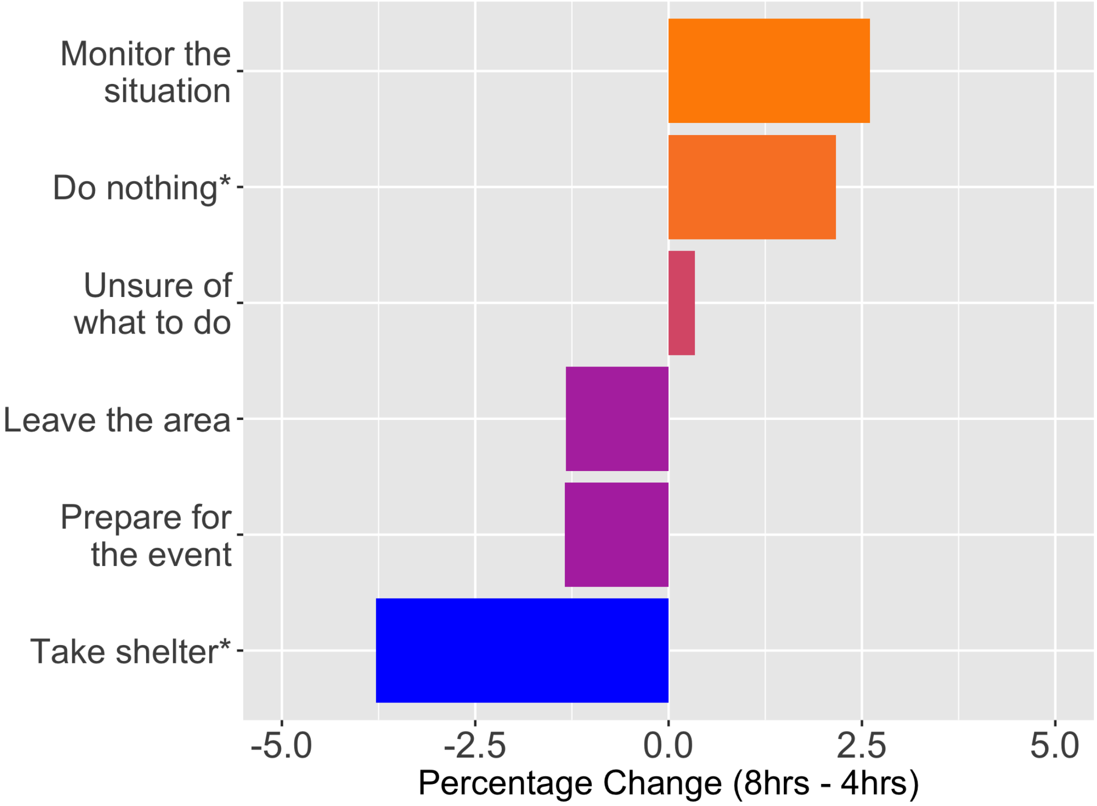


Figure 4.4: The difference in percentage of response categories between 8 hours (orange bars indicate more responses in this category) and 4 hours (blue bars indicate more responses in this category) of advance notice. * $p < 0.1$.

The difference in percentage of responses in each category reflects a shift from action to monitoring when shifting from 4 to 8 hours of advance notice (Fig. 4.4). The monitor, nothing, and unsure categories were more prevalent in the 8-hour group, although only the difference in the nothing category was statistically significant. Still, these changes may indicate that either 8 hours is too much advance notice, or that people are unaware of what actions they can or should be doing with an entire work day's worth of time. The 4-hour group is more focused on sheltering and preparing (with differences of 1.3% and 3.8%, respectively), although the prepare category is well represented in both groups. The differences in the percentage of responses within the shelter category and the do nothing category were statistically significant, which may further indicate that respondents within the 4-hour category are more focused on sheltering than those in the 8-hour category.

4.4 Summary and conclusions

Recent NOAA initiatives like the Warn-On-Forecast and the FACETs projects have begun to usher in a glimpse of what forecast information could look like in the future (Rothfusz et al. 2018). Given that most severe weather reports at any location are confined to sub-daily time periods (Krocak and Brooks 2020), it is within the realm of possibility that forecasters may soon be able to give hours of notice for severe weather events. While some work has been done to begin understanding how the public will react to increased specificity in products with warning-scale lead times, little work has been done to show how hours of advance notice for these events will impact response actions. This is vitally important as any actions taken a few minutes before the event are dependent on the actions taken previously.

After fielding a national survey of 3,000 US adults, we analyze and categorize text responses based on their content. First and foremost, we find that response actions are largely the same, regardless of how much time people are given. Analysis of single words show that sheltering behaviors may be slightly more common with 4 hours of advance notice and monitoring behaviors may be slightly more common with 8 hours of advance notice. However, many nuances are lost when we just look at single words. Categorical analysis of the responses show preparation and monitoring were the most common behaviors, regardless of how much time respondents were given. Although small, the differences we do find focus on preparing the most valuable items and sheltering when given 4 hours of notice, and on monitoring the weather and confirming information as well as preparing home items, pets, and family members when given 8 hours of notice.

Perhaps more importantly, we find more uncertainty about what to do with 8 hours of advance notice than with 4 hours, which may indicate that either 8 hours is too much time before the event occurs or that many respondents do not have a well conceptualized list of the kinds actions they could take to prepare for severe weather with more time. It is important to recognize that respondents in our survey were likely working with knowledge of the current system to help them visualize what they would do in a completely different system. While some people may know their routine when given 15 minutes of lead time, they may have never thought about all of the additional actions they may want to take given hours of advance notice. When the respondents are stratified by region, we do see a slightly higher proportion of those in less tornado-prone areas (the eastern and western regions of the US) stating that they were unsure of what they should do in both time categories. Within the 4-hour category, the eastern and western regions show 2.4% of responses in the unsure category while the central and southern regions show 1.6% in the same category (the same percentages in the 8-hour category are 2.8% for the eastern and western regions vs. 1.9% for the central and southern regions). We find a similar result when the data is stratified by education level. Those with less education (i.e. a high school degree or lower) said they were unsure of what to do more often than those with more education (3.2% vs. 1.4% in the 4-hour category and 3.0% vs. 2.0% in the 8-hour category, respectively). Given that education and prior experience may help residents understand what actions to take to prepare for these events, we follow recent reports from the National Oceanic and Atmospheric Administration and the National Academies of Sciences in emphasizing the need for collaborative work between the physical and social sciences in the weather enterprise (NOAA 2015, National Academies of Sciences et al. 2018). We believe that implementing changes in product structure must coincide with an education or information campaign that explains the nature of the change and how residents can utilize that change to enhance their safety and resilience. As related to this work, we believe an education campaign should include information on some of the kinds of actions that people *can and should* take multiple hours before a tornado occurs to make sure that they are safe if (when) the storm hits.

We also recognize the limitations of this work, which leaves room for future projects and research paths. First and foremost, we focus on anticipated actions to a hypothetical event, which may differ from actual responses to a real event. Studies of actual behavior after tornadoes are needed to understand if and how intended actions differ from actual responses. Second, we study intended responses to a single hazard; tornadoes. While there is likely some overlap in preparatory actions, many of the relevant response actions for other weather hazards would likely be different, meaning the results of this study are not likely to be generalizable to other categories of weather hazards. Additionally, our survey data is collected using an online platform, meaning vulnerable populations (like the elderly or those living in poverty) are likely to be underrepresented. We therefore see a need to employ multiple collection methods, including interviews and focus groups that target these populations to ensure results are generalizable. Finally, we would again like to emphasize the need for accompanying education campaigns, which suggests a close relationship between researchers, forecasters, emergency responders and communities will be needed if a new system is to be implemented. We hope that this work begins the process of understanding if and how response actions may change with more notice, and where we should be investing time and money in education campaigns as the forecast system continues to evolve.

Chapter 4 items of note:

- Residents may focus slightly more on preparation actions with 8 hours of advanced notice as compared to 4 hours.
- Respondents are likely anchoring to the current system when thinking about the actions they would take in response to 4 or 8 hours of advanced notice for a tornado.
- There is more uncertainty regarding appropriate actions to take for 8 hours of advanced notice.
- Any changes to the current severe weather notification system need to coincide with appropriate recommendations for actions that residents should take.

Chapter 5

Discussion

As research scientists, we are often taught (and expected) to go into a project with a carefully laid plan. Milestone A will be achieved by date X, milestone B will be achieved by date Y, and so on. Unfortunately (or maybe fortunately), very few projects actually follow such schedules, particularly when they involve people (whom you actually have to get permission to gather data from). Science should be flexible, allowing researchers to follow the interesting questions or explore particularly challenging hiccups. As the world works through an unprecedented global pandemic, I can only wonder how research milestones will need to be altered, how we must be flexible with what we expect of each other because ultimately all research projects involve people (even if you are completing purely physical science work).

When meteorologists are forecasting severe weather, it is vitally important to remember that the actions someone takes hours and days before a severe weather event fundamentally impact how and when they can respond to a tornado warning. Most of the research that focuses on warning response tends to treat those actions in isolation, assuming that everyone has the same baseline ability to act. Unfortunately, this is far from reality for many people with dependent family members, disabilities, or a lack of resources. Certain populations simply cannot take adequate protective action with only 15 minutes of advanced warning. Therefore, in addition to trying to extend tornado warning lead times, we should also be striving to provide useful, actionable information on longer scales, like hours or even days before the event happens.

When you examine the SPC convective outlook product, it becomes clear that the timing of the event is a vital piece of missing information that is not *explicitly* provided (although many of the discussion sections of the convective outlook include some type of timing information). The first step to assessing *why* this is the case was to understand if it is physically possible to provide timing information. Daily severe weather reports were divided into convective outlook days (12 UTC to 12 UTC) and then the distribution of those reports were analyzed. Results indicate that a majority of severe weather events occur in a much smaller timeframe than 24 hours. In fact, greater than 95% of severe weather reports are captured in a 4-hour period of the day. Essentially, this means that forecasters could, in theory, identify a 4-hour time frame for each point on the convective outlook forecast when a majority of the reports would occur, and the forecasted probabilities would not change.

Given the knowledge that forecasters can potentially provide timing information within the convective outlook (such that residents could prepare for a shorter threat timeframe hours before the threat occurs), we wanted to assess how useful this information may be. After some initial trial and error with forecasters, it became clear that we needed to include users in this assessment. Emergency managers overwhelmingly find timing information to be a critical tool to make decisions ahead of an event. They also expressed the need for this information to be widely accessible and consistently produced. Forecasters recognize this need, but are still concerned about the accuracy of the information that can be reasonably produced on a daily basis. There is still more work to be done, but it is also important to recognize when the science has met the acceptable threshold to be valuable for users.

In addition to testing this information with creators and decision makers, it is also important to allow residents to explore the utility of new products or information. Using a nationally representative survey of US adults, we show that resident response to a forecasted tornado hours in the future is not much different from traditional tornado warning response. We postulate that this may be due to anchoring behaviors to the current system. Further work should investigate how preparation actions may change specifically with vulnerable populations, although research has shown that it is exceedingly difficult to reach these populations.

Ultimately, this dissertation is an attempt to involve researchers, practitioners, and end users in the process of developing new forecast information. This co-production of scientific knowledge will only become more important as the National Weather Service moves towards more decision support services. Without intimate knowledge of how decision makers do their jobs, it is unclear how forecasters can produce actionable weather information. A formal development, testing, refinement, and implementation process also becomes more important as NOAA implements a more continuous flow of weather information, potentially unraveling some of the formal product structure that users have traditionally relied on to make decisions (like cancelling activities at the issuance of a tornado watch, for example). These changes *cannot* happen in a vacuum of researchers and forecasters if the goal is to have a seamless transition and magically more efficacious decisions. Such changes must ultimately be vetted with the people and organizations that rely on the information we produce. As scientists in a field that is connected to the daily lives of people arguably more than any other science field, it is our duty to ensure that we are serving the needs of residents at every step of our research process. To do that, scientists, funding agencies, and policy makers must be flexible to enough to allow researchers to see the testing and refinements of scientific knowledge as an asset to the process, not an obligation.

Chapter 5 items of note:

- Science is an inherently human process.
- Timing information for severe weather is important to many user groups.
- The testing of potentially operational products *must* be done with both creators and users.
- The co-production of scientific information will become more important as the weather enterprise moves towards more decision support services.
- The scientific process must be flexible enough to allow the testing and refinement of scientific information to be seen as progress, not setbacks or obligation.

Bibliography

American Meteorological Society Council, 2008: Enhancing weather information with probability forecasts. *Bulletin of the American Meteorological Society*, **89**, 1049–1053.

Armitage, C. J., and M. Conner, 2001: Efficacy of the theory of planned behaviour: A meta-analytic review. *British journal of social psychology*, **40 (4)**, 471–499.

Ashley, W. S., 2007: Spatial and temporal analysis of tornado fatalities in the united states: 1880–2005. *Weather and Forecasting*, **22 (6)**, 1214–1228.

Balluz, L., L. Schieve, T. Holmes, S. Kiezak, and J. Malilay, 2000: Predictors for people's response to a tornado warning: Arkansas, 1 march 1997. *Disasters*, **24 (1)**, 71–77.

Baumgart, L. A., E. J. Bass, B. Philips, and K. Kloesel, 2008: Emergency management decision making during severe weather. *Weather and Forecasting*, **23 (6)**, 1268–1279.

Brooks, H. E., C. A. Doswell III, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the united states. *Weather and Forecasting*, **18 (4)**, 626–640.

Cash, D. W., J. C. Borck, and A. G. Patt, 2006: Countering the loading-dock approach to linking science and decision making: comparative analysis of el niño/southern oscillation (enso) forecasting systems. *Science, technology, & human values*, **31 (4)**, 465–494.

Casteel, M. A., 2016: Communicating increased risk: An empirical investigation of the national weather service's impact-based warnings. *Weather, Climate, and Society*, **8 (3)**, 219–232.

Casteel, M. A., 2018: An empirical assessment of impact based tornado warnings on shelter in place decisions. *International journal of disaster risk reduction*, **30**, 25–33.

Chaney, P. L., G. S. Weaver, S. A. Youngblood, and K. Pitts, 2013: Household preparedness for tornado hazards: The 2011 disaster in dekalb county, alabama. *Weather, Climate, and Society*, **5 (4)**, 345–358.

Clark, A., and Coauthors, 2016: Spring forecasting experiment 2016 program overview and operations plan.

Clark, A., and Coauthors, 2017: Spring forecasting experiment 2017 program overview and operations plan.

Clark, A., and Coauthors, 2019: Spring forecasting experiment 2019 program overview and operations plan.

Cornell, S., and Coauthors, 2013: Opening up knowledge systems for better responses to global environmental change. *Environmental Science & Policy*, **28**, 60–70.

Cross, R., D. Ladue, T. Kloss, and S. Ernst, 2019: When uncertainty is certain: The creation and effects of amiable distrust between emergency managers and forecast information in the southeastern united states. *14th Symp. on Societal Applications*.

Demuth, J. L., R. E. Morss, B. H. Morrow, and J. K. Lazo, 2012: Creation and communication of hurricane risk information. *Bulletin of the American Meteorological Society*, **93 (8)**, 1133–1145.

Demuth, J. L., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for nws forecasters. *Weather and Forecasting*, **(2020)**.

Doswell, C. A., and D. W. Burgess, 1988: On some issues of united states tornado climatology. *Monthly Weather Review*, **116 (2)**, 495–501.

Ewald, R., and J. L. Guyer, 2002: The ideal lead time for tornado warnings-a look from the customer's perspective. *Publications, Agencies and Staff of the US Department of Commerce*, 39.

Fundel, V. J., N. Fleischhut, S. M. Herzog, M. Göber, and R. Hagedorn, 2019: Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users. *Quarterly Journal of the Royal Meteorological Society*, **145**, 210–231.

Gallo, B., and Coauthors, 2018: Spring forecasting experiment 2018 program overview and operations plan.

Greenwood, D. J., and M. Levin, 2006: *Introduction to action research: Social research for social change*. SAGE publications.

Gutter, B. F., K. Sherman-Morris, and M. E. Brown, 2018: Severe weather watches and risk perception in a hypothetical decision experiment. *Weather, climate, and society*, **10 (4)**, 613–623.

Hales Jr, J. E., 1989: The crucial role of tornado watches in the issuance of warnings for significant tornadoes. *Natl. Wea. Dig*, **15 (4)**, 30–36.

Hoekstra, S., K. Klockow, R. Riley, J. Brotzge, H. Brooks, and S. Erickson, 2011: A preliminary look at the social perspective of warn-on-forecast: Preferred tornado warning lead time and the general public's perceptions of weather risks. *weather, climate, and society*, **3 (2)**, 128–140.

Jauernic, S. T., and M. S. Van Den Broeke, 2016: Perceptions of tornadoes, tornado risk, and tornado safety actions and their effects on warning response among nebraska undergraduates. *Natural Hazards*, **80 (1)**, 329–350.

Joslyn, S., and S. Savelli, 2010: Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, **17 (2)**, 180–195.

Kang, J. E., M. K. Lindell, and C. S. Prater, 2007: Hurricane evacuation expectations and actual behavior in hurricane lili 1. *Journal of Applied Social Psychology*, **37 (4)**, 887–903.

Kloprogge, P., and J. P. Van Der Sluijs, 2006: The inclusion of stakeholder knowledge and perspectives in integrated assessment of climate change. *Climatic Change*, **75 (3)**, 359–389.

Krocak, M. J., and H. E. Brooks, 2018: Climatological estimates of hourly tornado probability for the united states. *Weather and Forecasting*, **33 (1)**, 59–69.

Krocak, M. J., and H. E. Brooks, 2020: An analysis of subdaily severe thunderstorm probabilities for the united states. *Weather and Forecasting*, **35**, 107–122.

Lanicci, J. M., and T. T. Warner, 1991: A synoptic climatology of the elevated mixed-layer inversion over the southern great plains in spring. part i: Structure, dynamics, and seasonal evolution. *Weather and forecasting*, **6 (2)**, 181–197.

Lemos, M. C., and B. J. Morehouse, 2005: The co-production of science and policy in integrated climate assessments. *Global environmental change*, **15 (1)**, 57–68.

Lindell, M. K., S.-K. Huang, H.-L. Wei, and C. D. Samuelson, 2016: Perceptions and expected immediate reactions to tornado warning polygons. *Natural hazards*, **80 (1)**, 683–707.

Ling, C., L. Hua, C. D. Karstens, G. J. Stumpf, T. M. Smith, K. M. Kuhlman, and L. Rothfusz, 2015: A comparison between warngen system and probabilistic hazard information system for severe weather forecasting. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, Vol. 59, 1791–1795.

Mauser, W., G. Klepper, M. Rice, B. S. Schmalzbauer, H. Hackmann, R. Leemans, and H. Moore, 2013: Transdisciplinary global change research: the co-creation of knowledge for sustainability. *Current Opinion in Environmental Sustainability*, **5 (3-4)**, 420–431.

Meadow, A. M., D. B. Ferguson, Z. Guido, A. Horangic, G. Owen, and T. Wall, 2015: Moving toward the deliberate coproduction of climate science knowledge. *Weather, Climate, and Society*, **7 (2)**, 179–191.

Miran, S. M., C. Ling, A. Gerard, and L. Rothfusz, 2018a: The effect of providing probabilistic information about a tornado threat on people's protective actions. *Natural Hazards*, **94 (2)**, 743–758.

Miran, S. M., C. Ling, and L. Rothfusz, 2018b: Factors influencing people's decision-making during three consecutive tornado events. *International journal of disaster risk reduction*, **28**, 150–157.

Morrow, B. H., J. K. Lazo, J. Rhome, and J. Feyen, 2015: Improving storm surge risk communication: Stakeholder perspectives. *Bulletin of the American Meteorological Society*, **96 (1)**, 35–48.

Morss, R. E., J. L. Demuth, and J. K. Lazo, 2019: Communicating uncertainty in weather forecasts: A survey of the us public. *Weather and forecasting*, **34 (2)**.

Murphy, A. H., 1993: What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and forecasting*, **8 (2)**, 281–293.

National Academies of Sciences, E., Medicine, and Coauthors, 2018: *Integrating social and behavioral sciences within the weather enterprise*. National Academies Press.

National Research Council, 2006: *Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts*. National Academies Press.

NOAA, 2015: Vision and Strategy: Supporting NOAA's Mission with Social Science. https://www.performance.noaa.gov/wp-content/uploads/SSVS$_{F}inal_{0}73115.pdf$.

NOAA National Centers for Environmental Information, 2020: U.S. Billion-Dollar Weather and Climate Disasters. Accessed 16 March 2020, https://www.ncdc.noaa.gov/billions.

NOAA National Weather Service, 2019: National Weather Service mission statement. Accessed 29 April 2019, https://www.nws.noaa.gov/mission.php.

NOAA Storm Prediction Center, 2019a: SPC Products Page. Accessed 12 September 2019, https://www.spc.noaa.gov/misc/about.html.

NOAA Storm Prediction Center, 2019b: WCM Page. Accessed 12 September 2019, https://www.spc.noaa.gov/wcm/.

Pielke, R. A., 1997: Asking the right questions: Atmospheric sciences research and societal needs. *Bulletin of the American Meteorological Society*, **78 (2)**, 255–264.

Reed, J. R., and J. C. Senkbeil, 2019: Perception and comprehension of the extended forecast graphic: A survey of broadcast meteorologists and the public. *Bulletin of the American Meteorological Society*, **(2019)**.

Ripberger, J. T., C. L. Silva, H. C. Jenkins-Smith, and M. James, 2015: The influence of consequence-based messages on public responses to tornado warnings. *Bulletin of the American Meteorological Society*, **96 (4)**, 577–590.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: Facets: A proposed next-generation paradigm for high-impact weather forecasting. *Bulletin of the American Meteorological Society*, **99 (10)**, 2025–2043.

Savelli, S., and S. Joslyn, 2012: Boater safety: Communicating weather forecast information to high-stakes end users. *Weather, Climate, and Society*, **4 (1)**, 7–19.

Schultz, D. M., E. C. Gruntfest, M. H. Hayden, C. C. Benight, S. Drobot, and L. R. Barnes, 2010: Decision making by austin, texas, residents in hypothetical tornado scenarios. *Weather, Climate, and Society*, **2 (3)**, 249–254.

Silva, J. T. J.-S. H. C. K. M., C. L.; Ripberger, 2017: Establishing a baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous united states. http://risk.ou.edu/downloads/news/WX17-Reference-Report.pdf.

Silva, J. T. J.-S. H. C. K. M. W. W. W., C. L.; Ripberger, 2018: Continuing and refining the baseline: Public reception, understanding, and responses to severe weather forecasts and warnings in the contiguous united states. http://risk.ou.edu/downloads/news/WX18-Reference-Report.pdf.

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

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

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Weather and forecasting*, **21 (3)**, 408–415.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the us tornado database: 1954–2003. *Weather and Forecasting*, **21 (1)**, 86–93.

Wilson, K. A., P. L. Heinselman, P. S. Skinner, J. J. Choate, and K. E. Klockow-McClain, 2019: Meteorologists' interpretations of storm-scale ensemble-based forecast guidance. *Weather, Climate, and Society*, **11 (2)**, 337–354.

Appendix A

Survey instruments

A.1 Sample SFE forecaster survey questions

A.1.1 Daily forecast evaluation

pst_rate_d1: Please rate the performance of yesterday's PST areas from the lead forecaster.
numeric, 1–10

group_pst_rate: Please rate the performance of yesterday's PST areas from your forecasting group.
numeric, 1–10

pst_difficulty: How difficult was it to create the PST areas yesterday?
1 – Very difficult
2 – Difficult
3 – Neither difficult nor easy
4 – Easy
5 – Very easy

pst_challenge: What was the biggest challenge associated with creating PST areas yesterday?
[VERBATIM]

pst_guidance: How much do you agree or disagree with the following statement?
I felt I had sufficient observational and guidance information to make the PST forecast

yesterday.

1 – Strongly disagree

2 – Disagree

3 – Neither disagree nor agree

4 – Agree

5 – Strongly agree

first_guess_useful: Was the “first guess PST” guidance product from your ensemble subset useful?

1 – Yes

2 – No

first_guess_acc: Please describe the accuracy or inaccuracy of the “first guess PST” guidance product from your ensemble subset (i.e. what did it do well, what issues were there, etc.)

[VERBATIM]

pst_other_guidance: Are there any other guidance products you wish you had access to? (This could be products that currently exist or information that you wish was provided. Please provide examples).

[VERBATIM]

pst_improve: Is there anything else (like model fields, drawing capabilities, more time, etc.) that you need to improve your forecast? (Please provide examples).

[VERBATIM]

A.1.2 End of week evaluation

Isochrones:

iso_challenge: What is the biggest challenge when drawing the isochrones?

iso_forecaster_improve: How would you improve this product for the forecaster?

iso_users: How do you think users would interpret their severe threat at a given point between two lines?

iso_users_improve: How would you improve this product for the user?

Potential Severe Timing Areas:

Thank you for participating in the 2019 Spring Forecasting Experiment. One of the projects you interacted with was the Potential Severe Timing (PST) product. As a reminder, this product is meant to accompany the SPC Convective outlook and represents forecasts of severe weather timing. Each area represents a 4-hour timeframe when severe weather reports are most likely to occur. The time labelled with each area is the start of the 4-hour period.

overlap_choice: Below is an example of a forecasted PST area. Please indicate which timeframe is forecasted for the red star.

22 - 02 UTC

00 - 04 UTC

22 - 04 UTC

23 - 03 UTC

Other **overlap_text** [VERBATIM]

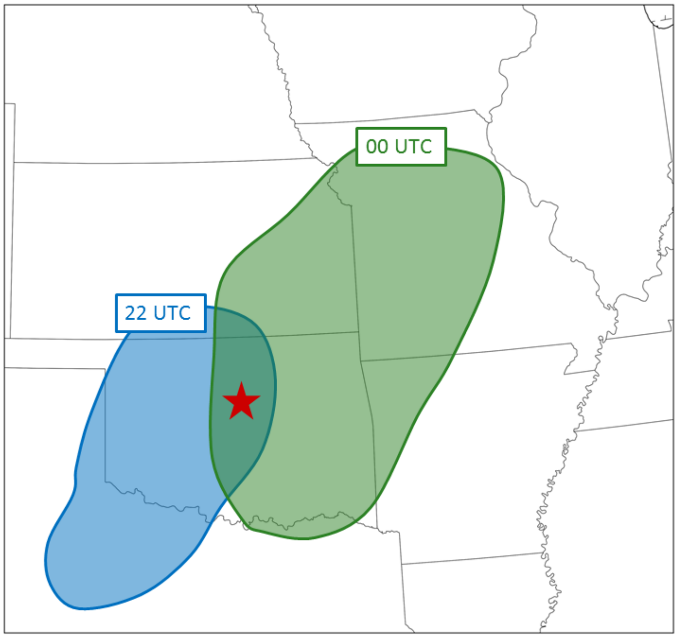


Figure A.1: An example PST graphic.

overlap_necc: Do you believe that overlapping PST areas are necessary?

1 – Yes, almost all of the time

2 – Yes, in some situations (please give an example) **overlap_necc_text** [VERBATIM]

3 – No, there should only be one period forecasted for each location

pst_aud_choice: Who do you think the audience for this product should be? (check all that apply).

1 – NWS WFO Forecasters

2 – Partners (like emergency managers, broadcast meteorologists, etc.)

3 – The general public

4 – None, it doesn't seem useful

5 – Other: **pst_aud_text** [VERBATIM]

pst_value: How much do you agree or disagree with the following statement?

The added value of the PST product is greater than the added workload.

1 – Strongly disagree

2 – Disagree

3 – Neither disagree nor agree

4 – Agree

5 – Strongly agree

A.2 Sample end user survey and focus group questions

A.2.1 Pre-test

Name: Upon being selected to participate in this year's experiment, you were asked to give a pseudonym that you would use throughout the study. Please type your pseudonym in the box below:
[VERBATIM]

wx_info_seek: In general, do you actively seek weather information as part of your job duties?

1 – Yes

2 – No

wx_info_sent: Is weather information send to you as part of your organization's standard operating procedures?

1 – Yes

2 – No

prod_used: What official products do you usually look at for weather information? (check all that apply)

1 – Storm Prediction Center convective outlook

2 – Storm Prediction Center mesoscale discussions

3 – Storm Prediction Center watches

4 – local NWS office briefings

5 – broadcast meteorologist posts/forecasts

6 – private firm forecasts/apps

7 – other **prod_used_text** [VERBATIM]

wx_info_qual: On severe weather days, what qualities of weather information are most useful to you before the event starts for your area? (select all that apply) 1 – Primary Hazards type

2 – Intensity

3 – Location

4 – Timing

5 – Other **wx_info_qual_text** [VERBATIM]

pre_storm_tasks: Think about your daily tasks on a severe weather day. What decisions need to be made or what tasks are you required to complete before watches are issued for your area? Please give examples of what the decisions/tasks are and when those decisions/tasks must be completed.
[VERBATIM]

pre_storm_tasks_prod: What forecast products do you use to help make those decisions or accomplish those tasks?

1 – Storm Prediction Center convective outlook

2 – Storm Prediction Center mesoscale discussions

3 – Storm Prediction Center watches

4 – local NWS office briefings

5 – broadcast meteorologist posts/forecasts

6 – private firm forecasts/apps

7 – other **pre_storm_tasks_prod_text** [VERBATIM]

watch_act: If a severe thunderstorm watch or tornado watch is issued for your area, are there additional decisions or tasks that must be completed prior to warnings being issued for your area?

1 – If yes, please describe: **watch_act_text** [VERBATIM]

2 – No

prim_haz_task: Does your procedure change depending on what the primary hazard for the day is (i.e. if tornadoes are the main threat vs. wind/hail as the main threat)?

1 – If yes, please explain how your procedure changes: **prim_haz_task_text** [VERBATIM]

2 – No

gend Gender:

1 – Male

2 – Female

3 – Other (please specify): **gend_text** [VERBATIM]

Age: What is your age:

[VERBATIM]

edu: Highest education you have attained:

1 – I completed some high school, but did not graduate

2 – High school diploma or equivalent (e.g., GED)

3 – I completed some college, but did not graduate

4 – Associate's Degree

5 – Bachelor's Degree

6 – Master's Degree

7 – Professional Degree (e.g., JD, MD)

8 – Doctoral Degree (e.g., PhD)

A.2.2 Mid-experiment micro-survey

Name: Pseudonym you've used this week:
[VERBATIM]

pst_missing_info: Is there information missing from the PST product that you would have liked to have had in preparation for this event?
Yes, please explain: **pst_missing_info_text** [VERBATIM]
No

pst_extra_info: Is there information you do not need that was included in the PST product?
Yes, please explain: **pst_extra_info_text** [VERBATIM]
No

A.2.3 Post-test

Name: Pseudonym you've used this week:
[VERBATIM]

useful_SPC_conv: In general, how useful were the following tools at informing your decisions this week? - SPC Convective Outlooks

1 – Extremely useless

2 – Moderately useless

3 – Slightly useless

4 – Neither useful nor useless

5 – Slightly useful

6 – Moderately useful

7 – Extremely useful

useful_PST: In general, how useful were the following tools at informing your decisions this week? – Potential Severe Timing – Mostly likely 4-hour window

1 – Extremely useless

2 – Moderately useless

3 – Slightly useless

4 – Neither useful nor useless

5 – Slightly useful

6 – Moderately useful

7 – Extremely useful

pst_ease: Please evaluate the degree to which the PST: - Was easy to use

1 – Strongly disagree

2 – Disagree

3 – Somewhat disagree

4 – Neither agree nor disagree

5 – Somewhat agree

6 – Agree

7 – Strongly agree

pst_info: Please evaluate the degree to which the PST: - Delivered pertinent information

1 – Strongly disagree

2 – Disagree

3 – Somewhat disagree

4 – Neither agree nor disagree

5 – Somewhat agree

6 – Agree

7 – Strongly agree

pst_quickly: Please evaluate the degree to which the PST: - Delivered information quickly

1 – Strongly disagree

2 – Disagree

3 – Somewhat disagree

4 – Neither agree nor disagree

5 – Somewhat agree

6 – Agree

7 – Strongly agree

pst_conf_dec: Please evaluate the degree to which the PST: - Made you more confident in your decisions

1 – Strongly disagree

2 – Disagree

3 – Somewhat disagree

4 – Neither agree nor disagree

5 – Somewhat agree

6 – Agree

7 – Strongly agree

haz_type_dec: In the days leading up to a potential severe weather event, do your decisions change if the main threat is wind/hail instead of tornadoes? If so, please offer a brief (2-3 sentence) description of the different decisions you need to make.

[VERBATIM]

pst_length: In the PST product, we have shown you a 4-hour window of severe weather threat. Would your decisions change if you are given a 6- or 8-hour window of severe threat?

Yes (please explain) **pst_length_text** [VERBATIM]

No

pst_task_change: Would a product like the PST or moving probabilities change the order of your task/decisions on a typical severe weather day?

Yes (please explain) **pst_task_change_text** [VERBATIM]

No

pst_dec_earlier: Would a product like the PST or moving probabilities allow for some decisions to be made earlier in the day?

Yes, please give an example: **pst_dec_earlier_text** [VERBATIM]

No

pst_issuance: This week, we have shown you the PST product with the 11:30am CT (1630z) Day 1 SPC Convective Outlook. However, we are considering different times for when this product could be issued. If you had to choose, would you like this product:

1 – With the 11:30am CT (1630z) Convective Outlook as it was presented this week

2 – With the 8:00am CT (1300z) Convective Outlook

3 – Other time frame (please explain): **pst_issuance_text** [VERBATIM]

pst_issuance_why: Why is this your preference? [VERBATIM]

pst_interp: How difficult was it to interpret the PST areas and the time frame associated with them?

1 – Extremely Difficult

2 – Somewhat Difficult

3 – Neither easy nor difficult

4 – Somewhat easy

5 – Extremely easy

pst_help_aid: How much do you agree or disagree with the following statement? The PST product would help aid in my decision making process during severe weather days.

1 – Strongly disagree

2 – Somewhat disagree

3 – Neither agree nor disagree

4 – Somewhat agree

5 – Strongly agree

pst_used_often: How much do you agree or disagree with the following statement? I believe I would use the PST on most severe weather days for my area.

1 – Strongly disagree

2 – Somewhat disagree

3 – Neither agree nor disagree

4 – Somewhat agree

5 – Strongly agree

pst_added_value: How much do you agree or disagree with the following statement? The added value of the PST product is greater than the added workload for forecasters.

1 – Strongly disagree

2 – Somewhat disagree

3 – Neither agree nor disagree

4 – Somewhat agree

5 – Strongly agree

pst_overlap: Below is an example of a forecasted PST area. Please indicate which time-frame is forecasted for the red star.

5:00pm – 9:00pm

7:00pm – 11:00pm

5:00pm – 11:00pm

6:00pm – 10:00pm

Other: **pst_overlap_text** [VERBATIM]

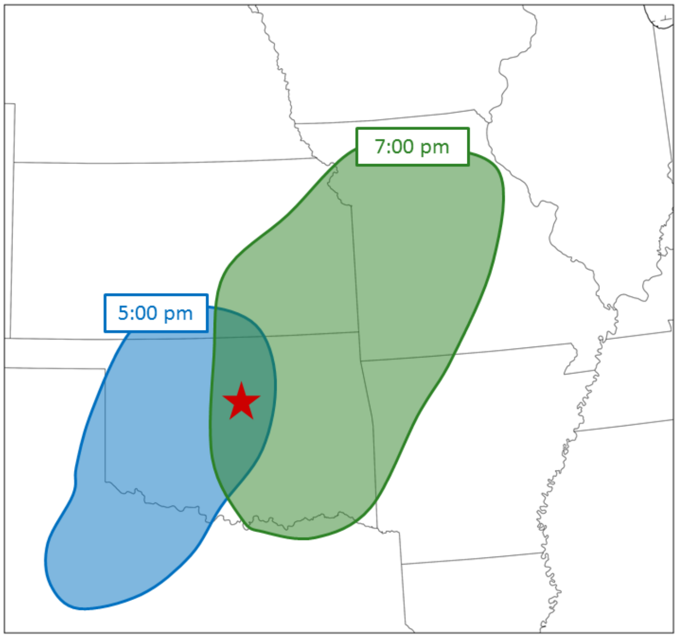


Figure A.2: An example PST graphic.

pst_overlap_useful: Do you believe that overlapping PST areas are useful?

1 – Yes, almost all of the time

2 – Yes, in some situations (please give an example) **pst_overlap_useful_text** [VERBATIM]

3 – No, there should only be one period forecasted for each location

A.2.4 EM focus group questions

Day 2 – 4

Tell us about how you interpreted the products as time went on, for example, are you looking at trends?

If so, what trends are you noticing?

What information did you find most helpful (e.g. overall likelihood for the day, evolution of event through the day, timing, hazard type, etc.)?

Is there any information that you want to have that is not provided?

What decisions did you make in this time period and why?

Day 1 operational products

Tell us about how you interpreted the products as time went on.

Are you looking at trends?

If so, what trends are you noticing?

What information did you find most helpful (e.g. overall likelihood for the day, evolution of event through the day, timing, hazard type, etc.)?

Is there any information that you want to have that is not provided?

What decisions did you make in this time period and why?

Now, you've seen three Day 1 products. How likely are you to look at this product at its issue time?

Does this depend on the last Day 2 update?

PST product

Tell us about your interpretation of these new products.

Are you noticing any trends?

What information is most important to you now?

What other information (if any) would be helpful to you at this point in the event?

How was the PST helpful or unhelpful for your operations on a day like this? In what way?

Appendix B

Emergency Manager Participants

Table B.1: 2018 emergency manager participants

Number	**Jurisdiction type**	**State**
1	Healthcare system	Minnesota
2	City	Texas
3	State	Georgia
4	County	Wyoming
5	State school system	Georgia
6	Multi-county	Missouri
7	County	Iowa
8	County	Georgia

Table B.2: 2019 emergency manager participants

Number	**Jurisdiction type**	**State**
9	State	Oklahoma
10	State	Ohio
11	County	Kentucky
12	City	Oklahoma
13	Hospital system	New York
14	State	Florida
15	Utility system	Colorado
16	County	Kentucky
17	City	Oklahoma
18	State	Colorado
19	County	Minnesota
20	County	Kansas
21	City	Georgia
22	County	Wisconsin
23	County	Iowa
24	Federal	Missouri
25	City	New York
26	City	Illinois
27	City	Oklahoma