UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

FORECASTER WARNING DECISION MAKING WITH RAPIDLY-UPDATING RADAR DATA

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

KATIE ANN WILSON
Norman, Oklahoma
2017

FORECASTER WARNING DECISION MAKING WITH
RAPIDLY-UDPATING RADAR DATA

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

Dr. David B. Parsons, Chair

Dr. Pamela L. Heinselman

Dr. Robert D. Palmer

Dr. Phillip B. Chilson

Dr. Ziho Kang

Dedication

To my parents, for without your support, I could not have chosen this path.

Acknowledgements

The journey to achieving my PhD has been a truly enjoyable and fulfilling experience, and for that I thank my advisor, Dr. Pamela Heinselman. It goes without saying that Pam has been a model advisor since day one. She has supported my growth as a scientist by encouraging independent thought, skill development, and proficiency in oral and written communication, while providing opportunities that have challenged my abilities and given me "real world" experience. I am also in awe of how well Pam has mastered being an advisor while simultaneously being a teacher, life-coach, and friend. Her professionalism, coupled with genuine concern and interest, has given me a sense of ease and stability throughout my time in graduate school. Thank you, Pam, for all that you have done for me.

I have also appreciated the support of my committee members throughout this process. I am thankful to Dr. Parsons for agreeing to chair this committee, to Dr. Palmer for patiently answering my weather radar questions and assisting me with scholarship applications, to Dr. Chilson for showing an interest in my research and inviting me to speak in his class, and to Dr. Kang for agreeing to collaborate with me and provide guidance on the eye-tracking research.

Many people contributed towards the work presented in this dissertation. Most importantly, this research would not have been possible without the participation of 31 National Weather Service forecasters, whose enthusiasm and dedication throughout the experiment did not go unnoticed. I will forever be grateful to these forecasters for all that I learned from them. Many thanks to Darrel Kingfield and Tiffany Meyer for providing technical support in the Hazardous Weather Testbed, to Gabe Garfield for creating the pre-briefing videos, to Greg Schoor for being a test subject, to James Murnan for his A/V assistance, and to Jeff Brogden, Chris Carter, and Robert Coggins for their IT support. I am also thankful to Charles Kuster for helping with data collection and analyses and for being a fantastic workmate throughout my entire time in graduate school. Finally, I appreciate the feedback from numerous colleagues and reviewers that helped to improve the writings presented in this dissertation.

It has been an honor to work alongside great people at the National Weather Center. The warm welcome from security staff, daily chit chat with office mates, and helpful assistance of friendly administrative staff has made coming to work each day a pleasure. External support from friends near and far has also helped me maintain a positive outlook on life during times that I felt overwhelmed and a little lost. Thank you, friends, for bearing with me as I embarked on this academic adventure.

Moving 5,000 miles away from my family to first participate in an exchange program and then return to graduate school was not an easy decision. It is a daily struggle to be so far from the people I love most in this world, but their encouragement, words of wisdom, and genuine excitement for the career I am pursuing remind me that the decisions I have made have been completely worth it. Like clockwork, Mum and Dad skype in each week to update me on life and to hear about my progress. There have been many times that, despite not truly understanding my research problems, their words have helped me most. My sisters, Vicky and Rachael, are two of my biggest cheerleaders. Their unconditional love, regardless of distance, has meant everything to me. Thanks for the Cadbury parcels, too- they have been pure medicine for my English soul!

Only one person has witnessed the up close, day-to-day highs and lows that I have experienced throughout graduate school. My husband, Chris, has been there for me every step of the way. He has celebrated my accomplishments with me, been a sounding board during times of confusion, and offered fresh perspectives when my goals felt foggy. I have always been able to depend on his love and support regardless of the stresses that his own job brings. Thank you, Chris. I am so glad to have had you by my side throughout this journey.

Funding for the research presented in this dissertation was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The contents of this dissertation do not necessarily reflect the views or official position of any organization of the U.S. government.

Table of Contents

Acknowledgments iv

List of Tables x

List of Figures xi

Abstract xv

Chapter 1: Introduction

1.1 Weather Warnings 1
1.2 Radar Observing Limitations 2
1.3 Integrating New Technology 5
1.4 The Phased Array Radar Innovative Sensing Experiment: Past Findings, Limitations, and Unanswered Questions 5
1.4.1 Warning Performance 6
1.4.2 Warning Decision Process 8
1.4.3 Forecasters' Feedback 11
1.5 Dissertation Outline 11

Chapter 2: Background

2.1 Radar 13
2.1.1 The Beginnings of Weather Radar 13
2.1.2 Advancements in Weather Radar 15
2.1.3 Temporal Sampling Limitations of the WSR-88D 17
2.1.4 Phased-Array Radar for Weather Observation 18
2.1.4.1 Technical Overview 18
2.1.4.2 Improving Scientific Understanding of Storm Processes 23
2.2 Studying Decision Making in Weather Forecasting 25
2.2.1 Learning through Surveys 25
2.2.2 Activities in the Testbed 26
2.2.3 Activities in a Naturalistic Setting 28
2.3 Measuring Mental Workload 30
2.3.1 Introduction to Mental Workload 30
2.3.2 Performance 32
2.3.3 Physiological Conditions 34
2.3.4 Subjective Ratings 36
2.3.5 Choice of Method 39
2.4 Eye Tracking 40
2.4.1 Early Discoveries of the Eye 40
2.4.2 Theoretical Advancements 43
2.4.3 Technology Advancements 44
2.4.4 Making Sense of Eye-Movement Data 49
2.4.5 Broadening the Research Applications 52

2.4.6 Eye Tracking for Meteorology 55

Chapter 3: Forecaster Performance and Workload: Does Radar Update Time Matter?

Abstract 58
3.1 Introduction 59
3.2 Methodology.. 62
3.2.1 Recruitment 63
3.2.2 Experimental Design 63
3.2.3 Methods 65
3.2.3.1 Working Events 65
3.2.3.2 Workload Ratings 66
3.3 Radar Data 67
3.4 Storm-Based Warning Verification 73
3.5 Performance 74
3.5.1 Performance: Severe Cases 78
3.5.2 Performance: Tornadic Cases 79
3.5.3 Performance: Null Cases 84
3.6 Warning Polygon Size and Duration 85
3.7 Cognitive Workload 89
3.7.1 Workload Distributions and Profiles 89
3.7.2 Reasoning for Higher Levels of Cognitive Workload 92
3.7.2.1 Categories 92
3.7.2.2 Temporal Resolution 94
3.7.2.3 Storm Type 96
3.8 Discussion 96
3.9 Conclusions and Future Work 99

Chapter 4: Exploring Applications of Eye Tracking in Operational Meteorology

Abstract 102
4.1 Background 103
4.2 Example: Understanding a Forecaster’s Decision Process 105
4.3 Counts and Durations of Eye Fixations 108
4.4 Fixation Trends 111
4.5 Future Applications 115

Chapter 5: Comparing Forecasters’ Eye Movements during the Warning Decision Process

Abstract 117
5.1 Introduction 118
5.2 Methodology 123
5.2.1 Experimental Design 123
5.2.2 Weather Scenario 124
5.2.3 Data Collection 126
5.2.4 Data Analysis 127
5.2.4.1 Fixation Identification 127

5.2.4.2 Areas of Interest and Fixation Measures 128
5.2.4.3 Scanpath Comparisons 130
5.2.4.4 Defined Stages 134
5.3 Results 136
5.3.1 First Five Minutes 136
5.3.2 Two Minutes Prior to Warning Decision 139
5.3.3 Warning Issuance Process 143
5.3.4 Warning Update Process 147
5.3.5 Differences in Duration 150
5.4 Discussion and Summary 150

Chapter 6: Considerations for Phased-Array Radar Data Use within the National Weather Service
Abstract 155
6.1 Introduction 156
6.1.1 Experiment Description 157
6.1.2 Focus Group Description 157
6.2 Using Rapidly-Updating Phased-Array Radar Data during the Experiment 159
6.2.1 Reactions to Radar Update Times 159
6.2.2 The Need to Adapt 161
6.2.3 Storm Trends 162
6.3 Integrating Rapidly-Updating Phased-Array Radar Data into Warning Operations 164
6.3.1 Visualizations 164
6.3.2 Training 165
6.3.3 Fatigue and Staffing 167
6.4 Discussion and Conclusions 167

Chapter 7: Conclusions, Implications, and Future Work 172

References 182

Appendix A: Focus Group Questions 204

Appendix B: List of Acronyms.. 205

List of Tables

Table 3.1. The Instantaneous Self-Assessment (ISA) tool adapted from Kirwan et al. (1997) 67

Table 3.2. Information about selected cases. Three case types are null, severe hail and/or wind, and tornadic 69

Table 3.3 The NWS event specific and generic verification systems used by the NWS and the hybrid verification system employed for PARISE 75

Table 3.4. The POD and FAR scores across all severe (SVR) and all tornado (TOR) warnings by update speed for a) all cases, b) severe cases Gamma, Beta, and Delta, and c) tornado cases Eta, Iota, and Zeta 76

Table 3.5. Overall median warning size and duration for all warnings issued in each case. Three case types are a) null, b) severe hail and/or wind, and c) tornadic 86

Table 3.6. Kruskal-Wallis rank-sum test and Wilcoxon-Mann-Whitney rank-sum test *p*-values for differences in cognitive workload distributions across groups with differing temporal resolution 92

List of Figures

Figure 1.1. The NOAA National Severe Storms Laboratory research PAR located at the National Weather Radar Testbed in Norman, Oklahoma (Photo courtesy of NOAA National Severe Storms Laboratory) 6

Figure 2.1. Photos from the Thunderstorm Project showing airplanes flying through thunderstorms, personnel operating a mobile radar unit, and the installation of a camera platform to observe cloud development. (Photos courtesy of NWS Wilmington, OH) 15

Figure 2.2. Weather and aircraft surveillance radar locations in the continental United States (Weber et al. 2007) 19

Figure 2.3. Installation of the PAR SPY-1A antenna at the National Weather Radar Testbed (Photo courtesy of A. Zahrai) 21

Figure 2.4. An example of a) the ADAPTS real-time display showing inactive (white), active (green), and neighboring (orange) beam positions (Heinselman and Torres 2011; Torres et al. 2012), along with an illustration comparing the locations sampled (red circles) using b) conventional scanning techniques and c) ADAPTS electronic scanning techniques for the same time period (Figure courtesy of Chris Curtis) 22

Figure 2.5. The relationship between (physiological) activation level, mental workload (task demands) and performance (taken from Young et al. 2015; originally adapted from De Waard 1996) 32

Figure 2.6. The relative density of cones (solid) and rods (dashed) and word identification accuracy (dotted) across the visual field (taken from Rayner et al. 2012) 43

Figure 2.7. Eye tracking setup using corneal reflections, a camera, and movement restraint (taken from Mackworth and Mackworth 1958) 45

Figure 2.8. Illustrations of eye gaze position determination using a) a head-mounted eye-tracking system and b) a remote-based eye-tracking system (Tobii 2017) 48

Figure 2.9. Front display of the Tobii TX300 system (Tobii 2014)............................ 49

Figure 3.1. Forecasters were recruited from the Great Plains region of the United States. The color bar indicates the number of forecasters participating from each of the eleven states 64

Figure 3.2. PAR 0.5° reflectivity for a) Epsilon, b) Alpha, and c) Theta (null tornado case). Green dots in c) are severe hail reports. Reflectivity (dBZ) color bar located at the top. White rings are displayed in 50-km increments.. 70

Figure 3.3. PAR 0.5° reflectivity for a) Beta, b) Gamma, and c) Delta. Green dots are severe hail reports and yellow dots are severe wind reports. Reflectivity (dBZ) color bar located at the top. White rings are displayed in 50-km increments............................ 71

Figure 3.4. PAR 0.5° reflectivity (left) and velocity (right) for a) Eta, b) Iota, and c) Zeta. Green dots are severe hail reports, and yellow dots are severe wind reports. Red dots are tornado reports (i.e., starting point of tornado path), while red lines in c) are tornado paths associated with longer-lived tornadoes. Thick white circles show location of couplet of interest. Reflectivity (dBZ) and velocity (m s-1) color bars located at the top. White rings are displayed in 50-km increments 72

Figure 3.5. Distribution of forecasters' severe warning lead times (min) for each case: a) Gamma, b) Beta, and c) Delta, organized by update speed. First, second, and third severe reports are denoted by numbers 1, 2, and 3 (magenta, blue, and red). For each report, the median severe warning lead time (min) for the full distribution is given by a dotted and annotated line (magenta, red, and blue)... 82

Figure 3.6. Distribution of forecasters' tornado warning lead times (min) for each case: a) Eta, b) Iota, and c) Zeta, organized by update speed. First, second, and third severe reports are denoted by numbers 1, 2, and 3 (magenta, blue, and red). For each report, the median severe warning lead time (min) for the full distribution is given by a dotted and annotated line (magenta, red, and blue)... 83

Figure 3.7. The False Alarm Ratio (FAR) for severe (black S) and tornado (red T) warnings by participant during Theta, plotted by update-speed group 85

Figure 3.8. Median severe thunderstorm warning a) size and b) duration for each participant group. Median values are included near the top of each bar. Radar-update speed (F=Full, H=Half, and Q=Quarter) worked by each group for each case is included near the bottom of each bar 88

Figure 3.9. Median tornado warning a) size and b) duration for each participant group. Median values are included near the top of each bar. Radar-update speed (F=Full, H=Half, and Q=Quarter) worked by each group for each case is included near the bottom of each bar 89

Figure 3.10. Boxplots of 5-min workload ratings for quarter-, half-, and full-speed groups for the null (a, b, c), severe (d, e, f), and tornadic (g, h, i) cases. The solid middle line indicates the median value and the box edges indicate the lower and upper quartiles (i.e., interquartile range). Minimum and maximum values are identified with whiskers and outliers are either less than 1.5 times the lower quartile or greater than 1.5 times the upper quartile.. 91

Figure 3.11. Reasoning categories for ISA ratings given at levels 4 and 5 for a) all groups combined, b) each temporal resolution, and c) each case type 95

Figure 4.1. Forecaster interrogating 1-min base velocity and reflectivity phased array radar data using the Warning Decision Support System-Integrated Information. The Tobii TX300 eye-tracker is positioned below the monitor 107

Figure 4.2. a) The Warning Decision Support System-Integrated Information display divided into three areas of interest: reflectivity (left panel, orange box), velocity (right panel, green box), and controls (bottom panel, blue box). Heatmaps were created for the b) total fixation count and c) absolute fixation duration for the entire case. Within the heatmaps, red values indicate a higher fixation count and absolute fixation duration, and blue colors indicate a lower fixation count and shorter absolute fixation duration..... 110

Figure 4.3. Boxplots showing the distribution of a) the 1-min fixation count and b) the 1-min mean fixation duration for the reflectivity, velocity, and controls AOIs. Boxplot whiskers indicate minimum and maximum values, the solid middle line indicates the median value, and lower and upper box edges indicate the interquartile range. Outliers are either less than 3/2 times the lower quartile or greater than 3/2 times the upper quartile. Strong evidence of differing medians is indicated by non-overlapping notches......... 111

Figure 4.4. a) Total fixation counts and (b) mean fixation durations within the reflectivity (orange), velocity (green), and controls (blue) AOIs per period................................ 112

Figure 4.5. A timeline of key observations made in the reflectivity (orange) and velocity (green) AOIs with respect to the northern storm (solid box) and southern storm (dashed box). Time is provided in the arrow for each period (top row) with corresponding case time (UTC) (bottom row). The timings of decisions to issue a warning are indicated by a red "w".. 114

Figure 5.1. Snapshot of the 0.5° reflectivity data at 2314 UTC 8 July 2014 125

Figure 5.2. Areas of interest are identified for the reflectivity data ("R", orange), velocity data ("V", green), control icons ("C", yellow), radar scan information ("S", grey), and the WarnGen interface ("W", blue). Note that the WarnGen interface appeared only when the forecaster selected to use it... 129

Figure 5.3. The a) five MultiMatch measures with corresponding examples of scanpaths that have b) relatively higher similarity scores and c) relatively lower similarity scores for two control participants. Adapted from Dewhurst et al. (2012)........................... 133

Figure 5.4. Examples of the four defined stages occurring within control participant C5's and experimental participant E6's warning decision process. With the exception of the first five minutes, the general duration of each stage lasted 1–3 minutes. Severe thunderstorm (SVR) issuance times are shown for the McClain storm (M in black) and Pontotoc storm (P in black), and first SVS issuance times for the McClain storm (m in red) and Pontotoc storm (p in red). Additional issuance decisions potentially impacting analyzed stages include issuance of a special weather statement (s in blue) and issuance of a second SVS for the McClain storm (m in blue). Vertical dashed lines indicate the timing of severe weather reports (see Fig. 5.5) 135

Figure 5.5. Control (top) and experimental (bottom) participants' warning products issued during the weather scenario. Markers are the same as in Figure 5.4. Vertical dashed lines indicate the timing of the severe weather reports associated with the McClain storm 136

Figure 5.6. Boxplot distributions of similarity scores for the five MultiMatch measures a) vector, b) direction, c) length, d) position, and e) duration for the control group (left position, black) and experimental group (right position, blue). Red boxes indicate distributions that are significantly different according to the Wilcoxon-Mann-Whitney rank-sum test (*p-value<0.05, **p-value<0.01, and ***p-value<0.001). Red crosses (+) indicate outlier values that are less (greater) than 1.5 times the lower (upper) quartile 138

Figure 5.7. Gaze plots depicting the scanpaths of participants a) C12, b) C15, and c) C10 in the two minutes prior to the Pontotoc storm warning decision. Circles represent fixations, the circle center identifies the fixation location, and the circle size characterizes the fixation duration. Lines between fixations represent the corresponding saccades. The background screenshot is the final frame from the period depicted........ 140

Figure 5.8. Gaze plots depicting the scanpaths of participants a) E11, b) E6, and c) E14 in the two minutes prior to the second McClain storm warning decision. Note that the WarnGen tool could be toggled on and off at any time and sometimes did not appear in the final screen capture........ 142

Figure 5.9. Gaze plots depicting the scanpaths of participants a) E2, b) E12, and c) E14 during the issuance of the first McClain storm warning 144

Figure 5.10. Gaze plots depicting the scanpaths of participants a) C4, b) C2, and c) C10 during the issuance of the second McClain storm warning........ 146

Figure 5.11. Gaze plots depicting the scanpaths of participants a) E10 and b) E5 during the first McClain storm warning update, and of participants c) C4 and d) C15 during the first Pontotoc storm warning update........ 149

Figure 6.1. Two groups of themes identified in transcriptions from forecasters' discussions during focus groups........ 159

Abstract

Phased-array radar is being considered as a potential future replacement technology for the current operational Weather Surveillance Radar 1988 Doppler system. One of the most notable differences in these weather radar systems is the temporal resolution. With phased-array radar collecting volumetric updates 4–6 times more frequently, the operational impacts of rapidly-updating radar data on forecasters' warning decision processes must be assessed. The Phased Array Radar Innovative Sensing Experiment (PARISE) was therefore designed to examine forecasters' warning performance and related warning decision processes during use of ~1-min radar updates in simulated real-time warning operation scenarios. While the 2010, 2012, and 2013 PARISE studies reported encouraging findings for forecasters' use of these data, each of these studies were limited in terms of sample size and the chosen methods. Additionally, important research questions that had not yet been explored remained unanswered. To address these limitations and investigate new research questions, thirty National Weather Service forecasters were invited to the NOAA Hazardous Weather Testbed to participate in the 2015 PARISE. Participating forecasters completed three components of this study: 1) the traditional experiment, 2) an eye-tracking experiment, and 3) a focus group.

The first component was designed to build on previous work by assessing and comparing forecasters' warning performance and related cognitive workload when using 1-min, 2-min, and 5-min phased-array radar updates during simulated warning operations. This traditional experiment was comprised of nine weather events that varied in terms of weather threat. Next, forecasters' eye movement data were observed as they each worked a single weather event with either 1-min or 5-min phased-array radar updates. This work was motivated by an eye-tracking pilot study, in which a forecaster’s eye movement data was found to correspond meaningfully to their retrospective recall data that described their warning decision process. The 2015 PARISE eye-tracking experiment allowed for an objective analysis of how forecasters interacted with a radar display and warning interface for a single weather event, and more specifically, supported an investigation of whether radar update speed impacts how forecasters distribute their attention. Lastly, six focus groups were conducted to enable forecasters to share their experiences on their use of rapidly-updating phased-array radar data during the experiment. The findings from the focus groups provide motivation for the integration of rapidly-updating radar data into the forecast office and highlight some important considerations for successful use of these data during warning operations. The work presented in this dissertation was approved by the University of Oklahoma’s Office of Human Research Participant Protection Institutional Review Board under projects #5226 and #5580.

Chapter 1

Introduction

1.1 Weather Warnings

Today, the official agency charged with issuing weather warnings in the United States is the National Weather Service (NWS). Comprised of 122 Weather Forecast Offices (WFOs), the NWS is responsible for collecting and processing billions of atmospheric observations and issuing approximately 1.5 million forecasts and 50,000 weather warnings each year (NOAA 2017a). Of these weather warnings, the past decade has seen the NWS issue a yearly average of 3,400 tornado warnings and 22,800 severe thunderstorm warnings (Harrison and Karstens 2017). Forecasters' decisions to issue weather warnings are based on their assessments of observations in real-time and their anticipation of severe weather in the near future (Brotzge and Donner 2013). A forecaster's attention during warning operations is therefore largely given to weather radar data because it provides observations of how storms are evolving in time and space. While interrogating radar data, forecasters apply conceptual models that are developed through education, training, and experience to interpret weather radar signatures and understand their importance.

Radar was first used for weather surveillance during World War II, and since, weather radar coverage across the United States has increased substantially and technological advancements have enhanced the observing capabilities of these systems. Notably, the installation of the Weather Surveillance Radar 1998 Doppler (WSR-88D) system contributed to past improvements in forecasters' abilities to detect severe weather hazards (Friday 1994). In 1986, prior to the installation of the WRS-88D, the average warning lead time for tornadoes was approximately 5 minutes and only 25% of tornado events were warned on (Brotzge and Donner 2013). After the WSR-88D network was installed, the proportion of tornadoes warned on increased to 75% and the average tornado warning lead time increased to 13 minutes (Erickson and Brooks 2006). These results, however, are influenced by the fact that the NWS assigns a warning lead time of zero minutes to unwarned tornado events. If these missed events are removed from the analysis, average tornado warning lead time remained steady throughout 1986–2004 at approximately 18.5 minutes (Erickson and Brooks 2006). This finding demonstrates that the longer average tornado warning lead time that followed the WSR-88D installation was due to fewer tornado events being missed and an improvement in the probability of detection.

1.2 Radar Observing Limitations

In more recent years, improvements in warning performance have plateaued and unwarned instances of severe weather hazards remain. In an effort to understand why these unwarned instances still occur, Quoetone et al. (2009) carried out a root cause analysis of 146 unwarned tornadoes that occurred during 2004–2009. Consulting NWS forecasters, Quoetone et al. (2009) found that in over two thirds of these cases, missed tornado events were due to radar-related issues such as sampling limitations and not detecting radar signatures indicative of tornadogenesis. Both NWS forecasters and broadcast meteorologists have also reported that unwarned tornado and severe weather events often occur due to insufficient information (LaDue et al. 2010). Specifically, the 4–6 minute volumetric update rate of the WSR-88D was reported as a limitation for observing storms transitioning into tornadic states. Furthermore, operational meteorologists reported that these temporal limitations can make interrogation of nontornadic severe thunderstorm threats challenging too. The onset of a downburst in which mid and upper-level precursor signatures are difficult to detect is just one example of this challenge (LaDue et al. 2010).

To address the temporal sampling limitations of the WSR-88D, new dynamic scanning methods have been developed and implemented into NWS operations, including Automated Volume Scan Evaluation and Termination (AVSET; Chrisman et al. 2009) and Supplemental Adaptive Intravolume Low-Level Scan (SAILS; Crum et al. 2013). The AVSET method terminates a volume scan and returns to the lowest elevation once the radar beam reaches the top of the precipitating cloud, meaning that storms that are shallower or farther from the radar will have faster volumetric updates. However, this method does not guarantee faster volumetric updates for storms that are deeper or closer to the radar. The SAILS method uses a scanning pattern that returns to the lowest level in the middle of the volume to provide one supplemental scan of the lowest elevation. More recently, the Multiple Elevation Scan Option-SAILS (MESO-SAILS; Chrisman 2014) was designed so that the operator can select two, three, or four supplemental scans of the lowest elevation during a volumetric update. The tradeoff for these more frequent low-level scans is an increase in overall volumetric update time, which in turn reduces the temporal sampling of mid and upper-level radar signatures.

An evaluation of forecasters’ use of AVSET, SAILS, and MESO-SAILS has not been completed following their implementation into NWS operations, and their impact on warning lead time is therefore unknown. However, given that these scanning methods cannot address radar temporal sampling issues for all weather scenarios, limitations of the WSR-88D system continue to hinder detection of rapidly-evolving severe weather. Furthermore, the initial improvement in warning performance owing to the implementation of the WSR-88D has not continued in recent years. This plateauing trend in lead time limits the NWS's ability to serve a rapidly growing and changing user community that, as described in a recent National Academy of Sciences (2012) report, expects "continuous improvement in public safety and property protection related to severe weather."

Addressing the limitations of the current radar system is one way to improve warning performance beyond today's capability. Given the age of the WSR-88D network and the lengthy process required for the development, testing, and deployment of a new system, considerations for a next generation radar network are already underway. Efforts have been largely focused towards phased-array radar (PAR) technology, which through electronic beam steering can scan the atmosphere with greater versatility than the WSR-88D. While the National Oceanic and Atmospheric Administration (NOAA) National Severe Storms Laboratory is investigating the feasibility of a multifunction S-band PAR network that will simultaneously meet both aircraft and weather surveillance needs (Zrnić et al. 2007; Stailey and Hondl 2016), the Engineering Research Center for Collaborative Adaptive Sensing of the Atmosphere is exploring the possibility of replacing the WSR-88D system with a dense network of ~10,000 X-band PARs that would be positioned on already-existing infrastructure across the contiguous United States (McLaughin et al. 2009).

1.3 Integrating New Technology

Upon review of the NWS Modernization and Associated Restructuring, the National Academy of Sciences (2012) outlined some of today's key challenges for providing outstanding weather service to the United States. One of these key challenges is to keep pace with quickly changing scientific and technological advancements. In response to this challenge, the National Academy of Sciences (2012) identified the need for operations-related research as a priority, with both research to operations and operations to research activities leading the way. This research, for example, would focus on the expected increase in the types and amounts of data that forecasters will receive (e.g., from radar, satellites, and numerical models) that need to be integrated effectively for successful communication of weather hazards. With PAR being a likely candidate for replacement of the current WSR-88D system, research to understand how rapid radar temporal sampling capabilities will impact forecasters' warning decision processes is essential. Forecasters will likely need to adapt how they process information and make warning decisions, and to guide this adaptation we must work to understand how forecasters will interact with the data.

1.4 The Phased Array Radar Innovative Sensing Experiment: Past Findings, Limitations, and Unanswered Questions

The NOAA National Severe Storms Laboratory's research PAR, which was located at the National Weather Radar Testbed in Norman, Oklahoma until May 2016 (Fig. 1.1) (Forsyth et al. 2005), has provided the opportunity to conduct behavioral research focused on NWS forecasters' use of PAR data. Loaned from the United States' Navy and adapted for weather use, this military PAR has collected radar data on numerous severe and tornadic thunderstorms in central Oklahoma. These archived weather cases have been used to assess impacts of rapidly-updating PAR data on NWS warning performance and related warning decision processes during the Phased Array Radar Innovative Sensing Experiment (PARISE).

Figure 1.1. The NOAA National Severe Storms Laboratory research PAR located at the National Weather Radar Testbed in Norman, Oklahoma (Photo courtesy of NOAA National Severe Storms Laboratory).

Prior to the work presented in this dissertation, the 2010, 2012, and 2013 PARISE studies investigated the impacts of rapidly-updating PAR data for specific weather scenarios. In each of these studies, twelve NWS forecasters were invited to visit the NOAA Hazardous Weather Testbed to participate in simplified warning operation tasks in simulated real-time. During these tasks, only reflectivity, velocity, and spectrum width PAR data were made available, and forecasters were asked to work these events like they would if they were in their WFOs. In each study, forecasters' warning lead time and verification statistics were calculated, and a variety of qualitative research methods were used to learn about their warning decision processes.

1.4.1 Warning Performance

Given that a known challenge within the NWS is providing warning lead time for weak and short-lived tornadoes, the first PARISE focused on paired forecasters' warning decisions for an event comprised of two supercells, of which one produced an EF1-rated tornado lasting only 3 minutes. This experiment found that forecasters using 43-s updates achieved longer tornado warning lead times than those using 4.5-min updates, but use of these 43-s data also resulted in a higher number of false alarms (Heinselman et al. 2012). The 2012 PARISE participants worked four weather events independently with 1-min PAR volumetric updates. Two of these events produced weak tornadoes, while the other two were null with respect to tornadoes and were chosen to further examine possible issuance of false alarm warnings. Forecasters' warning performance results during this study included a 20-min median tornado warning lead time (which exceeded the national average lead time for EF0/EF1 tornadoes by 7 min), and a probability of false alarm score better than chance (i.e., <0.5) for all but one forecaster (Heinselman et al. 2015). The efforts of PARISE were extended to severe hail and wind events in the 2015 PARISE, where forecasters' severe thunderstorm warning lead time was found to be statistically significantly longer during use of 1-min PAR updates compared to 5-min PAR updates (21.5 min vs. 17.3 min) (Bowden et al. 2015).

Improvements in forecasters' warning performance during use of rapidly-updating PAR data during the 2010, 2012, and 2013 PARISE studies are promising. However, to build on these earlier efforts, limitations in the chosen experimental designs of these past studies need to be addressed and research questions that have gone unanswered must be investigated. A limitation in each of these studies is sample size, both in terms of the number of participating forecasters and the number of cases that were worked. Additionally, the cases worked in each study focused on specific weather threats, which is unlike real-world operations where forecasters observe a variety of weather threats and storm types. To improve the generalizability of findings, the sample size of participants and the sample size of cases worked were increased in the traditional experiment component of the 2015 PARISE (Chapter 3; Wilson et al. 2017). The increase in the number of cases worked allowed for a variety of weather events to be introduced in the 2015 PARISE experimental design. In addition to the sample size limitation, prior studies only exposed forecasters to one of two temporal resolutions of radar data. However, forecasters have expressed interest in viewing 2-min PAR updates (Bowden and Heinselman 2016). Since forecasters' needs should drive radar requirements, use of this temporal resolution was also tested. Finally, the impact of rapidly-updating radar data on forecasters' cognitive workload is an important research topic that has not been previously considered. Addressing this topic is particularly important for ensuring that an increase in available radar data will not be detrimental to forecasters' warning performance and overall well-being. Therefore, the Instantaneous Self-Assessment tool was used to obtain forecasters' subjective ratings of their experienced cognitive workload.

1.4.2 Warning Decision Process

Learning about forecasters' warning decision processes has been a goal of PARISE from the very beginning. In the 2010 PARISE, audio and video recording of paired forecasters' activities and interactions captured the complex nature of their decision making due to different levels of experience, use of conceptual models, confidence, tolerance of missed events, perceived threats, and software issues (Heinselman et al. 2012). However, the accuracy of the observational data and the subjectivity inherent in the analysis process limited the reliability of the qualitative findings. Therefore, a cognitive task analysis method was applied in the 2012 and 2013

PARISE studies (Heinselman et al. 2015; Bowden and Heinselman 2016). Following the Recent Case Walkthrough procedure (Hoffman 2005), forecasters watched a playback video of their onscreen activity from a case they had just worked and recalled each minute what they were seeing, thinking and doing. An important finding from the recall data collected in the 2012 PARISE was that forecasters achieving above average tornado warning lead time applied conceptual models dependent on observations of mesocyclone trends seen in 1-min PAR update scans (Heinselman et al. 2015). Had these forecasters been using conventional ~5-min radar updates, these trends would have been difficult to observe and warning decision would likely have been delayed.

The recall data collected in the 2013 PARISE was analyzed within a situational awareness framework and thematically coded for perception, comprehension, and projection (Endsley 1995; Bowden and Heinselman 2016), and the frequency of these codes across the control (5-min PAR updates) and experimental (1-min PAR updates) groups was compared. While the groups did not differ in projection, the experimental group made statements of comprehension more often and recalled statistically significantly more perceptions than the control group (Bowden and Heinselman 2016). Given the superior warning performance of the experimental group in this study, we hypothesize that their higher number of perceptions resulted in the improved quality of their comprehensions and projections. This hypothesis is supported by evidence of experimental participants making more mastery decisions (i.e., confident and correct) than control participants during this study (Bowden et al. 2015).

Large amounts of qualitative data have been collected through use of the Recent Case Walkthrough procedure, and these data have provided new insight into forecasters’ warning decision processes during use of rapidly-updating radar data. Although these data have proved to be valuable for developing an understanding and appreciation of forecasters' approaches to warning operations, the chosen method has notable limitations related to both data collection and data analysis. First, forecasters' retrospective recalls are subject to inaccuracies, incompleteness, and biases. Second, making sense of the masses of qualitative data that forecasters provide is extremely challenging and labor intensive.

With these limitations in mind, a method to collect accurate and objective data on forecasters' cognitive processes during simulated real-time experiments was sought. Since many research domains (e.g., medicine and aviation) have successfully used eye tracking as a means for studying experts' cognitive processes during complex tasks, the possibility of using this method within the PARISE setting to better understand forecasters' cognitive processes through their distribution of visual attention was explored. A pilot study was first conducted to test whether a single NWS forecaster's eye movement data could be collected within the desired experimental set up and to examine whether the eye movement data makes sense given what we have already learned about forecasters' warning decision processes (Chapter 4; Wilson et al. 2016). Eye-tracking methods applied in the pilot study were then extended to a larger-scale experiment, where differences in forecasters' visual attention across a radar data display and warning interface were analyzed both in terms of individual differences and with respect to forecasters' use of 1-min vs. 5-min PAR updates (Chapter 5). Both retrospective recall and onscreen video data were also collected during the eye-tracking experiment to provide contextual understanding when interpreting forecasters' eye movement analyses.

1.4.3 Forecasters' Feedback

The impacts of rapidly-updating radar data on forecasters' warning decision processes have been studied extensively in prior studies. However, of these studies, only the 2013 PARISE obtained forecaster feedback following completion of all tasks, and this feedback was based on their use of a single temporal resolution of radar data (Bowden and Heinselman 2016). The 2015 PARISE was unique in that forecasters were exposed to 1-min, 2-min, and 5-min PAR updates for a variety of weather events, and were thus positioned to provide balanced feedback on these three temporal resolutions. Focus groups were therefore conducted during the 2015 PARISE to gather forecasters' reflections and opinions on their use of rapidly-updating PAR data for different types of weather events (Chapter 6). Findings from these focus groups also highlight what considerations and concerns forecasters have for use of these data in future warning operations, which help to inform future research, training, and implementation guidelines.

1.5 Dissertation Outline

This dissertation begins with a background section that covers topics including weather radar, studying decision making in weather forecasting, measuring mental workload, and eye-tracking research methods (Chapter 2). Since the latter two topics are unfamiliar to the everyday meteorologist, this background section is intended to prepare the reader for applications of these new concepts in later chapters. Chapter 3 reports on the traditional experiment component of the 2015 PARISE, which was designed to build directly on the efforts of the 2010, 2012, and 2013 studies. Initial exploration of eye-tracking research methods within meteorology is presented in Chapter 4, with the large-scale eye-tracking experiment of the 2015 PARISE following in Chapter 5. The final research portion of this dissertation describes the focus group component of the 2015 PARISE, which provides forecasters with an opportunity to voice their opinions as an end user, and in turn allows the operational community to inform research (Chapter 6). Chapter 7 brings together each of the research components of this dissertation to summarize the work that has been completed, present final conclusions, and identify unanswered questions and future research opportunities.

Chapter 2

Background

2.1 Radar

2.1.1 The Beginnings of Weather Radar

The history of weather radar is rooted in the World War II era, in which applications of radar were intended to serve the purpose of detecting enemy aircraft. With the development of the cavity magnetron, military personnel soon realized that in addition to observing aircraft, microwave radar (with S-band and X-band wavelength) could sense precipitation (Fletcher 1990; Atlas and Ulbrich 1990). Although weather echoes were considered a nuisance throughout much of the war, some efforts were made to understand how they related to atmospheric phenomena. The first known publication of this topic examined weather radar observations from 1942–1943 and qualitatively described different types of precipitation sampled as well as how echoes differed for S-band and X-band observations (Bent 1943). Furthermore, the Air Weather Service recognized the usefulness of radar for making flight decisions, and therefore began collecting routine radar observations to save time and money by cancelling or redirecting flights during hazardous weather conditions (Best 1973). Owing to the realized benefits of radar for tracking weather, training programs specializing in radar meteorology were developed and completed by approximately 7,000 officers at the Air Corps School and within universities (Byers 1970; Hitschfield 1986).

Interest in using radar for meteorological applications increased post-World War II, leading to the formation of several organized projects in the 1940s. The United States Air Force's All Weather Flying Division developed a weather radar program in 1945 to examine how airborne weather radar can be used to avoid hazardous weather (Metcalf and Glover 1990). In the next year, the Department of Meteorology at the Massachusetts Institute of Technology set out to learn more about radar for understanding the scattering nature of hydrometeors, how radar signatures relate to weather, and to develop knowledge of meteorological processes (Austin and Geotis 1990). Simultaneously, the University of Chicago coordinated the Thunderstorm Project to investigate the structure of thunderstorms following a number of weather-related aircraft accidents. The timing of this multiagency effort at the end of World War II was extremely beneficial to the project due to the increased availability of equipment and trained personnel (Fig. 2.1). A three-dimensional analysis of aircraft data led to the development of the three-stage (cumulus, mature, and dissipating) model of thunderstorms (Braham 1948; Byers and Braham 1949), which continues to be foundational to our understanding of thunderstorm lifecycles today. Additionally, with radar and airplanes being at the core of a weather research project for the first time, another important outcome of the Thunderstorm Project was the successful use of radar for observing dangerous portions of thunderstorms and being able to safely direct airplanes within the vicinity of them.

The first Weather Radar Conference was held at the Massachusetts Institute of Technology in March of 1947, marking the establishment of a weather radar community. At this same time, the Weather Bureau began to acquire and modify military radars for a Basic Weather Radar Network. In 1956, Congress agreed to fund 31 WSR-57s following the landfall of hurricanes along the United States east coast where radars were not located, eventually expanding to 66 locations (Whiton et al. 1998a). By 1974, radars were deployed to an additional 83 locations (Whiton et al. 1998a).

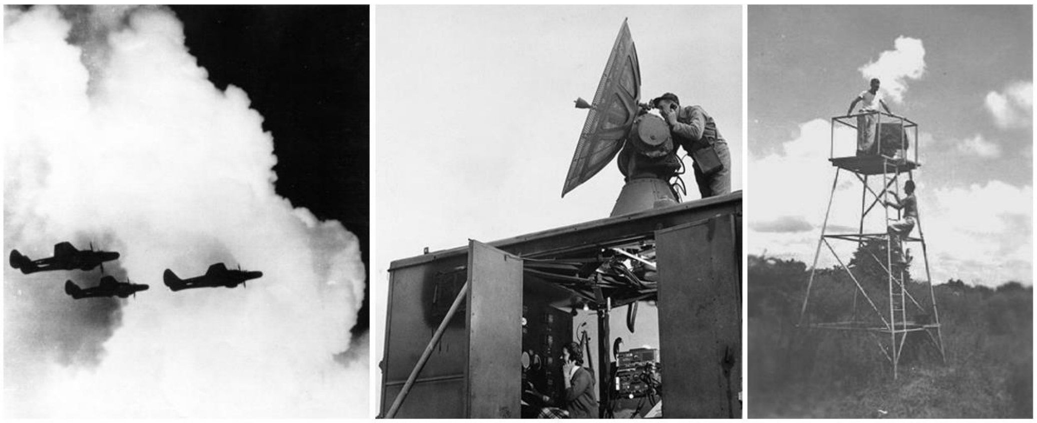


Figure 2.1. Photos from the Thunderstorm Project showing airplanes flying through thunderstorms, personnel operating a mobile radar unit, and the installation of a camera platform to observe cloud development. (Photos courtesy of NWS Wilmington, OH).

2.1.2 Advancements in Weather Radar

While reflectivity returns from the WSR-57s and WSR-74s provided data on storm structure and intensity, they did not provide information on storm motion. In turn, tracking radar signatures indicative of dangerous weather such as tornadoes and downbursts, as well as other meteorological phenomena, was not possible (Lemon et al. 1977; Brown et al. 1978; Wilson et al. 1980). While the application of the Doppler Effect to measure radial wind velocities was first proposed in the 1950s (Smith and Holmes 1961; Kessler 1990), it was not fully explored for operational purposes until the Joint Doppler Operation Project (JDOP) that took place during the spring seasons of 1977–1979 (Burgess et al. 1979; Brown and Lewis 2005). Improved detection of severe thunderstorms and increased tornado warning lead time were two important findings from JDOP that motivated the Doppler upgrade of the United States' national radar network (Atlas 1976). This network upgrade deployed 158 WSR-88Ds across the United States, forming the current operational next generation radar network (Whiton et al. 1998b). As discussed in the introduction, forecasters' improved ability to view radial velocity data positively impacted the detection of tornadogenesis and related tornado warnings. The WSR-88D has since continued to provide forecasters with a means to observe and interrogate thunderstorms, and has become an essential instrument to warning operations (Crum and Alberty 1993).

The most notable enhancement to the WSR-88D since its installment is the polarimetric upgrade (Istok et al. 2009). Up until recently, the single-polarization WSR-88D transmitted and received pulses of horizontally polarized electromagnetic radiation. Assessing the culmination of polarimetric weather radar research, a group of scientists and engineers recommended the need for an operationally-focused study to evaluate forecasters' use of polarimetric weather radar in real-time operations. The Joint Polarization Experiment (JPOLE) was therefore conducted to evaluate forecasters' use of polarimetric weather radar data for a variety of weather events (Ryzhkov et al. 2005; Scharfenberg et al. 2005). Forecasters reported that the polarimetric quantitative precipitation estimation algorithm was especially useful during rain events (Schuur et al. 2003). Furthermore, Zrnić and Ryzhkov (1999) identified the operational value of polarimetric weather radar data for improving rainfall estimation, hydrometeor classification, and discrimination of non-meteorological targets. These and other related studies motivated the polarimetric upgrade of the WSR-88Ds, and since 2013, these radars have transmitted and received both horizontally and vertically polarized electromagnetic waves. Both the amplitude and phase of signals returned in each polarization can now be compared to provide detailed information about the characteristics of targets in the atmosphere (Kumjian 2013).

2.1.3 Temporal Sampling Limitations of the WSR-88D

In a survey examining forecaster needs, the Radar Operations Center found that 62% of NWS forecasters felt they would benefit from faster-updating radar scans (Steadham 2008). Of these forecasters, 37% wanted these more frequent scans in the lower elevations, whereas 25% wanted these faster updates of the entire volume (Steadham 2008). As described in the introduction, efforts have been made to increase the frequency of radar data through the application of new scanning techniques such as AVSET, SAILS, and MESO-SAILS (Chrisman et al. 2009; Crum et al. 2013; Chrisman 2014). However, depending on the weather conditions and the distance of the storm from the radar, these new techniques do not always provide an ideal solution. Furthermore, while forecasters may benefit from the more frequent lower-level scans for monitoring tornado potential of storms, this improvement comes with a cost of slower overall volumetric updates. Unfortunately, increasing the rotation rate of the radar antenna to overcome the sampling trade-offs is not an option because of the detrimental impacts to data quality and the hardware of the system (Chrisman 2009). Therefore, the temporal sampling limitations of the WSR-88D is constrained to volumetric updates of 4–7 minutes during severe weather, which is known to hinder forecasters' abilities to detect the onset of weather threats including tornadoes and downbursts (Quoetone et al. 2009; LaDue et al. 2010).

The Doppler benefits of the WSR-88D for better detecting severe weather have been realized in operations, but in recent years improvements in warning performance have plateaued. Given the known temporal sampling limitations of the WSR-88D to warning operations, a next step to advancing warning lead time is to therefore address this technical constraint of weather radar. With this next step in mind, the future of the WSR-88D is under consideration, and while upgrades and maintenance keep these systems functioning beyond their expected 20-year lifetime, they will eventually need to be replaced (Saffle et al. 2009; Crum et al. 2013). Scientists have therefore been considering a future replacement technology to the WSR-88D, and the NOAA National Severe Storms Laboratory has identified PAR as a leading candidate (Weber et al. 2007; Zrnić et al. 2007).

2.1.4 Phased-Array Radar for Weather Observation

2.1.4.1 Technical Overview

The United States' Navy has successfully used PAR on their cruiser ships for missile defense and aircraft detection purposes since the mid-1970s (Dranidis 2003). The antenna of a PAR system consists of numerous transmit-receive elements that allow for electronic steering of the radar beam (Zrnić et al. 2007). By controlling the timing (and therefore phase) of pulses transmitted in each element, the radar beam can be repositioned to any chosen azimuth or elevation almost instantaneously. This design differs substantially to the WSR-88D, in which a parabolic dish antenna is used to form a beam of energy that is then transmitted into the atmosphere. Unlike PAR, the WSR-88D steers the radar beam mechanically through rotation of the antenna. Because of this mechanical dependence, to collect one volume scan the WSR-88D antenna must rotate fully through 360° for a sequence of predetermined elevations.

The non-contiguous and versatile scanning abilities of PAR means that this technology does not have the temporal sampling limitations of the WSR-88D, thus making PAR a promising replacement candidate for future weather radar. Also notable is the potential for PAR to serve as an observing instrument for multiple federal agencies within the United States. A multifunction PAR network has been proposed to combine the currently 510 government-owned weather and aircraft surveillance radars in the United States (Fig. 2.2) to 334 multifunction PAR systems (Weber et al. 2007). Given that PAR can concentrate data collection in areas that are of interest while also being able to reposition focus quickly, the observing needs of multiple agencies may be met simultaneously. The replacement of these multiple independent networks to a consolidated network would provide the required radar coverage for each agency's mission while also reducing the required training, maintenance, and operation costs.

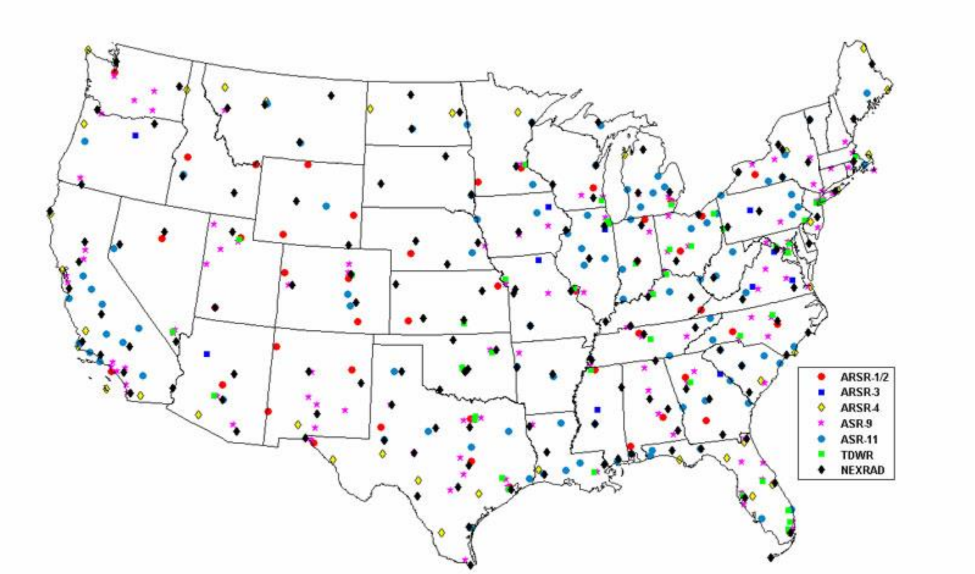


Figure 2.2. Weather and aircraft surveillance radar locations in the continental United States (Weber et al. 2007).

Scientists at the NOAA National Severe Storms Laboratory have collaborated with other government, academic, and private sector entities to examine the suitability of PAR technology specifically for weather observation. Important to this collaboration was the formation of the National Weather Radar Testbed in Norman, Oklahoma, which houses a phased array SPY-1A antenna loaned from the United States Navy (Fig. 2.3) (Forsyth et al. 2005). Since the original purpose of this military radar was to detect aircraft and missiles, it was first modified for weather observation before beginning data collection in the spring of 2004. Many characteristics of this research PAR are similar to the WSR-88D, in that it operates at S-band and with a comparable range and range resolution. However, whereas the WSR-88D operates with a 1° beamwidth, PAR operates with a non-conformal transmit beamwidth that gradually increases from 1.5° to 2.1° as the beam moves from boresight to ±45° (Zrnić et al. 2007). This difference is due to the flat-panel array design of the PAR.

The greatest difference between the WSR-88D and the PAR is that the former steers the beam mechanically while the latter uses its 4352 transmit-receive elements to steer the beam electronically. The electronic steering also has the advantage of removing beam smearing effects during data collection. Given that the PAR consists of one single panel that observes a 90° sector at one time, it takes less than a quarter of the time to obtain a volume update compared to the WSR-88D. To achieve this higher-temporal resolution for a full 360° coverage, a future operational PAR system of this design would be comprised of four flat-panel arrays that each observe a 90° sector (Brown and Wood 2012).

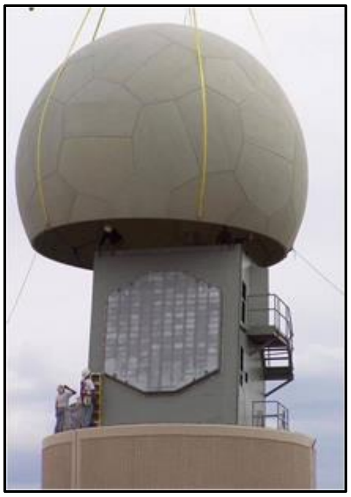


Figure 2.3. Installation of the PAR SPY-1A antenna at the National Weather Radar Testbed (Photo courtesy of A. Zahrai).

In addition to reducing the volumetric update time through sampling only a 90° sector, adaptive scanning strategies are employed to further control the temporal resolution of PAR. The Adaptive Digital Signal Processing Algorithm for PAR Timely Scans (ADAPTS) is used to trade spatial resolution and/or data quality to provide faster radar updates (Heinselman and Torres 2011). ADAPTS uses a criteria that determines whether beam positions should be active or inactive to enforce weather-focused scanning (Fig. 2.4a) (Heinselman and Torres 2011). The significance criteria for activating a beam position depends on whether reflectivity values have met a pre-defined threshold and whether these reflectivity values have sufficient spatial coverage. Next, beam positions that are within close proximity to those that meet the significance criteria are considered to have neighborhood significance and are also activated. Unlike conventional scanning methods where regions without weather are sampled (Fig. 2.4b), adaptive scanning methods allow for weather-focused observations as well as the sampling of other targets of interest such as aircraft (Fig. 2.4c).

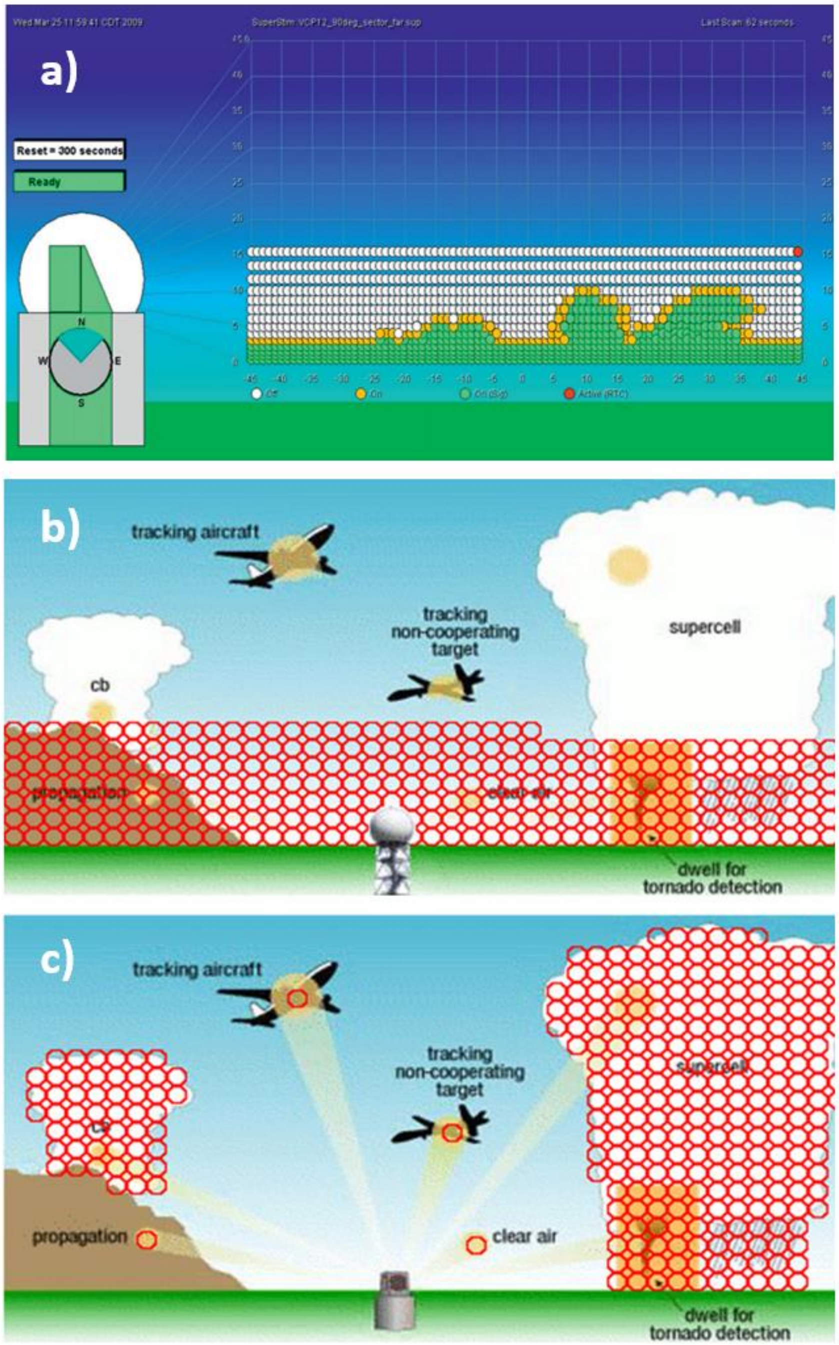


Figure 2.4. An example of a) the ADAPTS real-time display showing inactive (white), active (green), and neighboring (orange) beam positions (Heinselman and Torres 2011; Torres et al. 2012), along with an illustration comparing the locations sampled (red circles) using b) conventional scanning techniques and c) ADAPTS electronic scanning techniques for the same time period (Figure courtesy of Chris Curtis).

2.1.4.2 Improving Scientific Understanding of Storm Processes

Since the spring of 2004, the National Weather Radar Testbed PAR has collected data on a variety of weather events. Scientists have examined these rapidly-updating radar data to examine the finer temporal detail of weather phenomena and improve scientific understanding of storm processes. Heinselman et al. (2008) completed a first investigation into what adaptively-scanning higher-temporal resolution S-band radar data can observe compared to conventional WSR-88D data. Analysis of PAR and WSR-88D reflectivity and velocity data showed that the higher-temporal sampling of three convective storms allowed for a better depiction of their structures and evolutions. The velocity signatures of a reintensifying supercell were better captured, including the storm's inflow, convergence trends, and rotation, while the updraft development, descending high-reflectivity core, midlevel-altitude radial convergence, and low-level convergence associated with a microburst were observed successfully and with more temporal detail (Heinselman et al. 2008). Additionally, higher-temporal resolution reflectivity signatures associated with the reintensification of a hailstorm allowed for a more detailed analysis of its related storm structure, including the development of a bounded weak echo region, a high-reflectivity core, and a related three-body scatter spike (Heinselman et al. 2008).

Being able to observe storm features with improved temporal detail also aided in the identification of damaging wind mechanisms that were associated with a quasi-linear convective system (Newman and Heinselman 2012). Through the use of 1-min PAR updates, the evolution of mesovortex circulations, azimuthal shear, and descending reflectivity core were depicted more clearly, while the increased spatial resolution of

PAR’s vertical sampling proved useful for viewing the full structure of the midlevel jet (Newman and Heinselman 2012). Trends in 1-min PAR data have also proven important for capturing severe downburst precursor signatures that are evident only several minutes prior to downburst maximum intensity (Kuster et al. 2016). Furthermore, 26-s PAR updates have been used to compare reflectivity and velocity radar data to the lightning data of a hail storm (Emersic et al. 2011). This comparison was important for assessing how lightning activity relates to storm kinematics and related storm intensity (Emersic et al. 2011).

In addition to observational case study analyses, rapidly-updating PAR data have been used in numerical modeling studies. Tanamachi and Heinselman (2015) assimilated 1-min PAR updates into a numerical cloud model to create three-dimensional cloud-scale analyses. These analyses were used to better understand storm merger processes through objective identification of storm updrafts and vortices that were analyzed prior to, during, and after the storm merger event. Observing system simulation experiments have also demonstrated that the assimilation of rapidly-updating radar observations results in more realistic analyses and ensemble forecasts of convective storms than when conventional WSR-88D observations are used (Xue and Droegemeier 2006; Yussouf and Stensrud 2010). Furthermore, these more realistic depictions of storms can result in an improved alignment between the locations of high probability low-level vorticity with radar-derived storm rotation (Supinie et al. 2017).

Case study analyses have established that higher-temporal resolution radar data provide enhanced observations of storm features, structures, and evolutions that are otherwise unobservable in traditional WSR-88D data. These data help to identify the complex nature of storms and the dynamic interactions that occur within them. Given that forecasters have expressed a need to observe these temporal details during operations (Steadham 2008; LaDue et al. 2010), applications of these enhanced observations within the decision making environment is therefore also important to investigate.

2.2 Studying Decision Making in Weather Forecasting

2.2.1 Learning through Surveys

Forecasters provide both a crucial and complex human element to the process of weather forecasting. The dissemination of surveys have been useful for gathering information on the forecaster decision making process for many decades. A substantial advantage of this research method is its far reach; forecasters from all across the United States can contribute their perspectives at a relatively low cost to the research group. Early studies using this method began exploring some of the subjective aspects associated with forecasters' judgment calls. For example, a nationwide survey tested NWS forecasters on different aspects of precipitation probability forecasting (Murphy and Winkler 1971; Murphy and Winkler 1974). Findings from this survey recognized several challenges related to the subjective influences on probability-based judgments, such as forecasters' confusion over probabilistic concepts and their tendency to hedge when stating their degree of belief. The inherent uncertainty in meteorology means that forecasters' abilities to make informed assessments of it is essential to weather prediction. More recently, Novak et al. (2008) surveyed NWS forecasters on their assessments of uncertainty with use of various guidance and products, the training that is available to support these assessments, and what operational challenges forecasters currently face in expressing uncertainty information in forecasts. The finding that operational forecasters believe that they should play a significant role in communicating uncertainty, but that there is not yet agreement on the degree to which forecasters should modify objective ensemble guidance, demonstrates that the development of uncertainty information must be a collaborative effort between forecasters, model developers, and its users (Novak et al. 2008).

Abilities to communicate other weather hazards effectively have also been explored through the dissemination of surveys. Given the increasing cost associated with ice storm impacts, Call (2008) surveyed warning coordination meteorologists on their understanding of ice storm hazards, their related warning procedures, and how they communicate information about the hazard to members of the community. Additionally, with the projected frequency of extreme heat events expected to increase, Hawkins et al. (2017) completed an internal assessment with WFOs to document current decision making related to the issuance of heat-based products and to develop ideas for better communicating extreme heat risks. The various NWS-focused research questions that surveys have successfully answered demonstrate that they are an effective method for learning about the current state of forecasters' knowledge and procedures within the WFO. An important outcome of findings from these surveys is the generation of ideas that need to be explored further and the suggested recommendations that will support improvements to forecast operations.

2.2.2 Activities in the Testbed

The benefits of incorporating forecasters into the research and development process of new technologies and resulting data have been demonstrated in the JDOP and JPOLE studies discussed in the previous section. Taking this "end-to-end-to-end" approach (Morss et al. 2005), effective collaboration between researchers, forecasters, and software developers will result in the implementation of advancements that are both scientifically sound and operationally relevant. The emergence of numerous NOAA testbeds across the United States have made these collaborations possible (Ralph et al. 2013). The NOAA Hazardous Weather Testbed, located in Norman, Oklahoma focuses on severe weather prediction in a quasi-operational environment. This testbed is home to both the Experimental Forecast Program and the Experimental Warning Program. The Experimental Forecast Program hosts an annual Spring Forecasting Experiment that focuses on the use of numerical model guidance for producing outlook products beyond those that are currently issued operationally (e.g., Kain et al. 2003; Clark et al. 2012; Gallo et al. 2017). Some participants attending the Spring Forecasting Experiment are operational meteorologists, though the majority of participants are research scientists actively working in model development.

The Experimental Warning Program functions separately to the Experimental Forecast Program, and focuses on nowcasting capability and the warning decision process. The annual activities of the Experimental Warning Program are comprised of numerous projects each guided by separate research groups. These projects have focused on NWS forecasters' use of: numerical weather prediction analyses during severe thunderstorm and tornado events (Calhoun et al. 2014), a prototype probabilistic hazard information tool (Karstens et al. 2015, 2016), products developed from the new Multi-Radar Multi-Sensor system (Smith et al. 2016), and rapidly-updating satellite (Line et al. 2014) and PAR (e.g., Heinselman et al. 2015; Wilson et al. 2017) data during the warning decision process. Additionally, integrated warning teams consisting of NWS forecasters, broadcast meteorologists, and emergency managers have been studied in the testbed to better understand the interactions and relationships necessary for preparing for and responding to high-impact weather events successfully (e.g., LaDue et al. 2017; Obermeier et al. 2017).

While forecaster performance with respect to predictive skill is often assessed for testbed activities, a substantial effort has been made to ensure that new products, tools, and data are also evaluated from the perspectives of participating forecasters. Researchers conducting studies within the testbed have used a variety of qualitative methods to obtain data on forecasters' perspectives, including observations, group discussions, interviews, surveys, and blog posts. Cognitive task analysis methods have also been used to elicit detailed retrospective recalls of forecasters' warning decision processes during their use of rapidly-updating PAR data (Heinselman et al. 2015; Bowden et al. 2016). Additionally, human factors specialists have conducted usability studies of new decision-support systems for weather forecasters within the Hazardous Weather Testbed (Ling et al. 2015; Argyle et al. 2016). These studies have proven important for identifying usability issues and providing recommendations for improved operational meteorology software.

2.2.3 Activities in a Naturalistic Setting

Although testbeds are designed to simulate aspects of a real operational setting, the ability to control for external factors can limit the realism of the decision making environment. Therefore, researchers have also completed naturalistic studies within the WFO. Here, forecasters' cognitive processes and skills are applied to tasks in real time and to complex problems that have genuine consequences (Lipshitz 2001; Klein 2008; Gore et al. 2015). Studies within WFOs can be short and focused, or they may be lengthy and broad. For example, soon after the implementation of the WSR-88D at the Raleigh, North Carolina WFO, forecasters' warning decision processes were examined (Hoium et al. 1997). Researchers from North Carolina State University logged forecasters' use of reflectivity, velocity, and ground truth data for a variety of weather events. To assume a participant-observer role and to blend into the WFO, these researchers also contributed to routine tasks such as analyzing surface charts. Morss and Ralph (2007) also took a participant-observer role in their analysis of forecaster use of additional meteorological information during the California Land-falling Jets and Pacific Land-falling Jets experiments. Additional data including wind, melting level, surface, dropsonde, and radar observations were made available to WFOs located on the west coast of the United States in real time. These data were expected to aid forecast decisions during flooding and winter storm events. Morss and Ralph (2007) mainly used observation and semi-structured interviews to understand how forecasters used these additional information in their forecasts, but informal discussions and interviews were also carried out when possible.

Unlike the focused studies described above, Daipha (2015) completed an ethnography within a single WFO that required a multiple-year-long effort. Rather than studying one aspect of the forecaster decision making process, Daipha (2015) immersed herself into the complex system of a northeastern WFO, observing the office culture, use of ground truth, data, and technology, and chosen methods of communication. More recently, Henderson et al.'s (2017) ethnography has also considered some of the social, political, and ethical challenges that forecasters face during the warning decision process. Finally, while most ethnographic studies in WFOs have been contained to a single location, Friedman et al.'s (2015) ethnography involved visits to 11 WFOs, during which forecasters' use of social media within an uncertain decision making environment was observed and documented. Different to studies conducted through the means of surveys and testbeds, researchers' time spent in WFOs allow them to build meaningful relationships with forecasters and develop a deeper understanding of the interconnecting factors and nuisances governing forecasters' everyday work activity. Though completing studies in this manner can be time intensive and logistically challenging for researchers, the knowledge gained from these experiences can be invaluable.

2.3 Measuring Mental Workload

2.3.1 Introduction to Mental Workload

With the continued development and integration of new technology into the work place, mental workload is an ever growing topic of interest (Wickens and McCarley 2008). To optimize system performance in human-machine systems, the mental workload of an operator is an important consideration. Mental workload refers to the amount of attention resources required to meet the desired performance criteria of a system, and is influenced by task demands and past experience of the operator (Young and Stanton, 2005). In the past, much of the mental workload research has focused on transport-related systems, such as air traffic control, aviation, and especially driving (Da Silva 2014; Young et al. 2015). Applications of this research have also been useful within military and medical professions, and more recently for evaluating operator use of modern-day technology such as computers and smartphones (Hart 2006).

Assessing the cognitive demands of a system helps ensure both the well-being of the operator and that optimal system performance is achievable. If an operator's cognitive load is too low or too high, their performance can suffer, potentially resulting in undesirable outcomes (Cain 2007; Mehta and Parasuraman 2013). Figure 2.5 illustrates this relationship between mental workload, task demands, and performance (De Waard 1996; Young and Stanton 2015). As task demands increase, the operator's mental workload also increases. A corresponding improvement is observed in performance during this initial increase, which plateaus at an optimum level while task demands continue to climb and mental workload increases accordingly. However, when task demands begin to exceed the operator's available attention resources, mental workload becomes unmanageable, and performance consequently deteriorates. This state is referred to as overload, and occurs when the operator is unable to process all presented stimuli. In this instance, the operator can become distracted, and the use of selective attention to acquire and process information can be insufficient (Young and Stanton 2005; Young et al. 2015).

Poor performance can also be observed when mental workload levels are too low. This state is referred to as underload and occurs when the task does not provide enough stimulation to keep the operator engaged. Rather, a lack of stimuli results in lower levels of alertness and attention, and the operator thus lacks vigilance when monitoring the situation at hand (Young and Stanton 2005, 2015). A decoupling of performance and mental workload is therefore evident (Fig. 2.5), such that increases in mental workload is beneficial to performance when the task demands are lower and resources necessary to meet the increasing demand are available, but once the task demands exceed available resources, the increasing levels of mental workload become damaging to performance (Young et al. 2015).

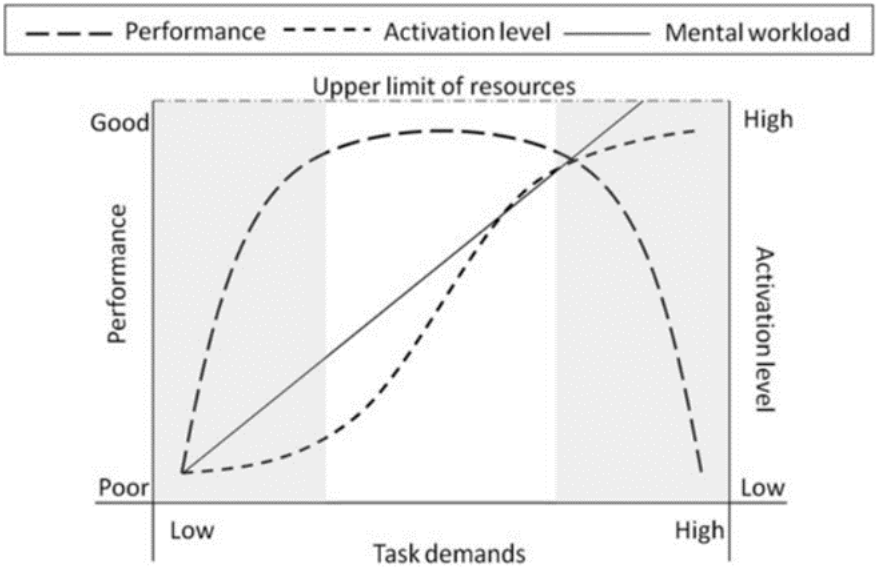


Figure 2.5. The relationship between (physiological) activation level, mental workload (task demands) and performance (taken from Young et al. 2015; originally adapted from De Waard 1996).

2.3.2 Performance

Researchers have developed a variety of methods to measure mental workload, and these can be classified into three types: 1) performance, 2) physiological conditions, and 3) subjective ratings. Performance can be analyzed using either the primary task or a secondary task. For both primary and secondary tasks, if an undesirable level of mental workload is imposed on an operator, their performance is expected to suffer (Proctor and Zandt 2008). Assessments of performance during the primary task is possible using metrics such as response time, accuracy, and root mean square error. The choice of metric depends on the nature of the primary task. For example, response time and accuracy would be suitable measures for a vigilance task, where an operator may be tasked with identifying a specific feature within a noisy image. The root mean square error might be useful in measuring the operator's deviation from the center lane position during a driving task. Performance of a secondary task is useful when trying to measure the operator's spare capacity. Compared to when the operator completes the primary and secondary tasks separately, the change in performance during the dual-task scenario indicates the additional cognitive demand on the operator and their related mental workload. The types of secondary tasks given to operators include mental math activities, estimations of elapsed time, and reaction times to other visual stimuli. For example, the peripheral detection task requires an operator to wear a headband that positions a light within their peripheral and acts as a secondary visual stimulus during a driving simulation (Schapp 2013). The operator is asked to press a button attached to their index finger whenever they see this light flash. Their response time and the number of missed flashes indicate the operator's spare visual attention resources and their associated mental workload. If the operator is overloaded, they will not be able to complete the secondary task successfully.

Degradations in performance for both the primary and secondary tasks indicate that the cognitive demands of the task exceeded the operator's available resources, and was thus overloaded. However, a limitation of this approach is that changes in mental workload are difficult to detect during times when performance is not impacted. The performance of an operator during a highly demanding task may be comparable to that of a task with lower demands if they are motivated and choose to exert greater effort. Individual motivation is an important factor in the observed dissociation between performance and subjective measures of mental workload (Vidulich and Wickens 1986; Yeh and Wickens 1988). An additional limitation of the secondary task is that it may be disruptive to the primary task, and practice of the dual-task scenario is required before stable performance is established (Proctor and Zandt 2008).

2.3.3 Physiological Conditions

With increased levels of workload, an operator’s brain activity and general level of arousal is expected to increase (Roscoe 1992; Proctor and Zandt 2008; Young et al. 2015). As a result, this increase in physiological activation can signal when suboptimal levels of workload are experienced. Unlike with performance or subjective rating methods, continuous monitoring of physiological conditions during a task is possible, meaning that an operator’s workload can be assessed on a much finer temporal scale, and transient fluctuations in workload that would usually go undetected can be observed (Mehta and Parasuraman 2013). Also, individual biases do not influence these measures of workload like they can do with subjective ratings.

A variety of methods have been used to study operators’ physiological conditions and associated workload. One example is the field of neuroergonomics, an interdisciplinary research approach that brings together neurology, ergonomics, and human factors to understand how the human brain functions within work settings as well as in natural settings (Parasuraman 2011; Mehta and Parasuraman 2013). Neuroimaging techniques, such as electroencephalography, is used to monitor the electrical activity of an operator’s brain during a task. Past studies have shown significant correlations between electroencephalography indices and cognitive states during tasks performed in real-time and in simulation mode (Wilson and Eggemeier 1991; Sterman and Mann 1995; Berka et al. 2004, 2007). These data can therefore be used to better understand the cognitive state and associated mental workload of an operator as they respond to stimuli.

Measurements of cardiovascular responses to stimuli, such as heart rate, heart rate variability, and blood pressure, have also proved useful for the study of workload. Roscoe (1992) describes early applications of this approach for evaluating first military pilots' and then civilian pilots' responses to stressful situations. Higher levels of mental workload have been related to increased heart rates, suppressed heart rate variability, and increased blood pressure (Roscoe 1992; Wilson 2002). Evidence of these physiological responses to higher levels of mental workload have also been observed on the ground. In high-traffic conditions, when air traffic controllers must be alert and monitoring the situation with high levels of attention, studies have shown that heart rate and blood pressure become statistically significantly higher (Vogt et al. 2006) and heart rate variability decreases (Hilburn 2003).

Other physiological observations that are indicators of mental workload include pupillary response, blinking activity, galvanic skin response, and cortisol levels. Eye-tracking methods are used to monitor pupillary response and blinking activity. Studies have shown that under conditions in which higher levels of mental workload are required, an operator's pupil dilates (Jorna 1997; Beatty 1982; Neumann and Lipp 2002; Hilburn 2003), and the time between two successive eye blinks increases while the duration of each eye blink decreases, especially for visually demanding tasks (Veltman and Gaillard 1998; Wilson 2002; Marquart et al. 2015). Galvanic skin response data provides a measure of skin conductance and thus psychological or physiological arousal, and have been shown to vary with changes in mental workload (Nourbakhsh et al. 2012). Finally, higher levels of mental workload can be exhibited in an operator's cortisol (also known as the "stress hormone") levels. Biochemical analysis of air traffic controllers' saliva samples showed increased cortisol levels during busier working periods in which mental workload was higher (Farmer et al. 1991). These results were also reflected in an everyday office-work environment (Cinaz et al. 2013).

Although physiological measurements can track the cognitive state of an operator in an objective manner and on a much finer temporal scale than performance measures or subjective ratings, this approach does have its limitations. One of the most notable limitations is the likely contamination of these physiological data. Contamination may come from the ambient environment (e.g., light sources will impact pupil size), or it may come from within the operator. Furthermore, an operator's body movements as well as their emotional states can confound these measures; it is difficult to separate external influences from the influence of workload itself. Additionally, many of these methods require careful calibration of instruments to individual operators, and the large signal to noise ratio often found in these data can make them difficult to analyze. Aside from these limitations, some of these methods require costly equipment and therefore may be unavailable to the researcher.

2.3.4 Subjective Ratings

An operator is able to provide their own valuable insight into their experienced mental workload that is unobtainable with alternative methods. Hart and Staveland (1988) suggested that "… subjective ratings may come closest to tapping the essence of mental workload." Many subjective rating tools have been designed for a range of purposes, though there is no agreement within the human factors community on which of these tools is best (Farmer and Brownson 2003). The complexity and intrusiveness of subjective workload scales varies, partly depending on whether they are multidimensional or unidimensional. Winter (2014) completed a literature search on workload to analyze which subjective workload rating tools have been used most frequently. The post popular tool was the multidiemsional NASA-Task Load Index (TLX), which evaluates an operator's mental demand, physical demand, temporal demand, performance, effort, and frustration levels on a continuous scale of 1–100, which together can be combined and weighted to give an overall workload level (Hart and Staveland 1988; Winter 2014). Winter (2014) argues that the popularity of the NASA-TLX does not necessarily stem from it being the best, but because it has become synonymous with workload due to the Matthew effect (Merton 1968). This effect is based on the idea that the "rich get richer and the poor get poorer"; popularity is gained through increased awareness which reinforces its use and eventually becomes the accepted standard. The comparable quality and sensitivity of other workload tools supports this argument, as well as the fact that limitations in the NASA-TLX design have not been addressed despite many years of use. These limitations include the anchor effect (i.e., participants tend to use only part of the scale), that the tool is being used differently across studies (e.g., some weight the subscales while others do not), and that no "redline" is defined for identifying when an operator's workload becomes too high (Hart 2006). Furthermore, the strong correlation between the subscales brings to question how well an operator is able to discriminate between the different types of workload described in this tool (Hart 2006).

Another multidimensional tool that has been used frequently is the Subjective Workload Assessment Tool (SWAT), which uses a card sorting procedure to allow operators to provide feedback on which tasks have higher demands (Reid and Nygren

1988). These cards describe three categories: time load (how limited time is and how many tasks must be completed), mental effort load (attentional demands of tasks), and psychological stress load (fatigue, emotional state, anxiety). Each category has three levels of intensity: low, medium, and high. Operators then do pairwise-comparisons of these 27 cards to rate overall mental workload (Luximon and Goonetilleke 2001). However, this multidimensional tool is complex and time intensive, and would certainly be intrusive if it were completed during a task. While the NASA-TLX has higher operator acceptance than the SWAT, it too can be time intensive.

Subjective rating tools with unidimensional scales have also been used in research studies. The modified Cooper-Harper scale was first designed to measure pilots' workload when handling aircraft (Cooper and Harper 1969), and since has been modified to suit other types of scenarios in which operators have to make decisions (Wierwille and Casali 1983). Operators follow a decision tree to determine their overall level of workload on a scale of 1–10, with 1 indicating that the task is easy to complete and the desired performance is easy to attain, and 10 indicating that the task is impossible and cannot be completed reliably (Proctor and Zandt 2008). Even simpler is the Instantaneous Self-Assessment (ISA) tool, which was designed to collect overall workload ratings quickly during a task (Jordan and Brennan 1992). Based on a rating scale of 1–5 (from "underutilized" to "excessive"), operators report how busy they are according to their perceived spare capacity. Ratings can be provided using a keypad specifically designed to collect ISA ratings (Hering and Coatleven 1996), or records can be kept more simply with a pen and paper. The ISA tool can be used to address workload in a variety of research areas, such as for assessing drivers' mental workload during different traffic volume conditions (Girard et al. 2005) and in driving scenarios when distractive thoughts are introduced (Lemercier et al. 2014). Sensitivity to changes in task demand, and thus mental workload, has been demonstrated in both these driving studies and when compared to physiological measures of workload (Tattersall and Foord 1996). Additionally, ISA ratings have shown to correlate well with NASA-TLX ratings as well as those from other workload tools (Farmer and Brownson 2003). However, an evaluation of the ISA tool found that primary-task performance decreased in conditions where ISA ratings were requested from operators during tasks (Tattersall and Foord 1996). Given that the ISA tool is considered one of the least intrusive subjective rating tools (Miller 2001; Farmer and Brownson 2003), the effect that more complex, multidimensional tools such as NASA-TLX and SWAT could have on primary-task performance is therefore concerning. This point reinforces the importance of choosing a tool that not only measures aspects of workload that are of interest to the researcher, but that also does not act to confound experimental data.

2.3.5 Choice of Method

When choosing a method for measuring mental workload, a number of considerations must be made (Miller 2001; Proctor and Zandt 2008). First, the method must be sensitive enough to detect changes in workload due to increased task demand. If the researcher wants to measure how much workload is imposed on different types of resources, then the method should also have good diagnostic skill. To ensure that different levels of workload are represented accurately and consistently, the method should provide valid and reliable data. The interval of data collection is also important; if observing transient fluctuations in workload is important to the research question, monitoring physiological conditions may be most appropriate. If overall workload is more useful to the research question, then performance measures or subjective rating tools capturing workload at the end of a task could be more suitable. However, Jansen et al. (2016) noted the importance of capturing changes in workload during a simulation rather than just at the end. In this instance, a simple unidimensional subjective workload rating tool may be preferred over the more time-intensive, multidimensional subjective workload tools. The intrusiveness of all methods should also be considered; the nature of intrusiveness will vary depending on the method, but may be in the form of disruption during a task (e.g., subjective workload ratings), or due to requirements for the operator to wear a monitoring device (e.g., physiological measures). This consideration relates to the importance of operator acceptance in the chosen method—researchers should ensure that operators are willing to use the chosen method in a correct manner. Additionally, the implementation of the chosen method to the overall experiment is key, such that it should practically make sense and support data collection rather than hinder it. For example, if a study wants to assess workload in a dynamic environment over a long period of time, it would not be possible for an operator to wear a head mounted eye-tracker or electrolyte sensors for this entire time. Finally, some of these methods require expensive equipment, most notably for measuring physiological conditions. The availability of such equipment may therefore limit the choice of method.

2.4 Eye Tracking

2.4.1 Early Discoveries of the Eye

Scientists first became intrigued by the role of eye movements during reading in the late 19th century and early 20th century (Jacob and Karn 2003). In this early pioneering work, a French scientist, Professor Emile Javal, reported that "…there is practically no reading, or rather no direct seeing of the words and letters, except during the pauses." From these observations, Javal was the first to report two basic eye movements that occur during reading: fixations and saccades (Huey 1908). Fixations occur when the eyes focus on a specific point and refer to the pauses that Javal observed. Although the eye appears still during fixations, slight movements still occur due to nystagmus and tendencies for the eye to drift away from and return to a point in very small and quick movements (Rayner 1998). Fixations generally last on the order of $250\ ms$, but reading studies have shown these durations to vary from as short as $50\ ms$ to as long as $600\ ms$ (Rayner 1998). Saccades, on the other hand, are much faster eye movements that can travel at a velocity of $500\ {}^\circ s^{-1}$ (Rayner 1998). They occur as the eye traverses between fixations, and are named after the French word for "jump." The rapidity of saccades result in saccadic suppression, meaning that a person is unable to acquire and process information adequately during these eye movements (Matin 1974).

The early discovery that eye movements relate to reading activity opened up a world in which human attention could be studied. The mental resource capacity of humans is limited, and attention is therefore used to direct resources to information that is most useful. Eye movements provide a representation of how visual attention is distributed (Duchowski 2007). While three regions of the eye characterize the visual field, it is within the foveal region that visual attention is greatest. This region extends up to 2° from the visual center and is made up of predominantly cone receptors (Fig. 2.6). These receptors allow details to be seen on a fine scale and make images appear more sharp and colorful, meaning that information placed within this region is viewed with high visual acuity (Rayner et al. 2012; Bojko 2013). The parafoveal region (2°–5°) extends farther out from the fovea and is where cone density decreases and rod density increases (Fig. 2.6). Rods are the dominant receptor in the peripheral region (beyond 5° from visual center) and are important for detecting motion and observing different levels of brightness (Rayner et al. 2012; Bojko 2013). Reading studies have therefore shown that the ability to discriminate text is best in the fovea region, and diminishes substantially as you move farther out in the visual field until it can no longer be read at all in the peripheral region (Fig. 2.6) (Rayner et al. 2012). Hence, it is fair to assume that our visual attention is given predominantly to the material we choose to fixate on within our fovea region.

Though the study of eye movements during reading emerged over 100 years ago, growth in this research area was slow. The psychology field moved away from cognitive research and focused more on behavioral observations. Acknowledging that cognitive processes cannot be directly observed, psychologists believed that the study of behaviorism would be a good approach for learning about language processing. This approach, however, did not come to fruition, and the importance of cognition to language processes, as well as many other mental processes, was accepted (Rayner et al. 2012). Interest in cognition was revived in the 1970s, which resulted in developments in both eye-tracking technology as well as theoretical understanding of the relationship between eye movements and cognitive processes (Poole and Ball 2006).

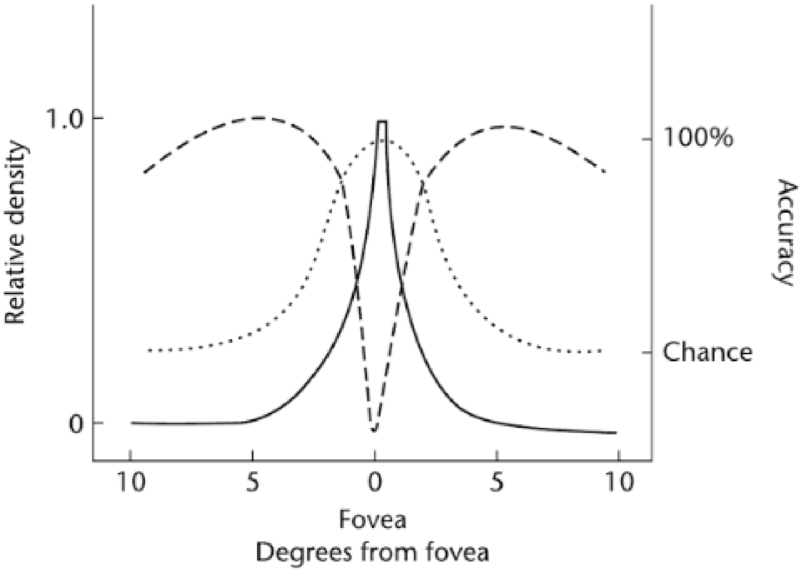


Figure 2.6. The relative density of cones (solid) and rods (dashed) and word identification accuracy (dotted) across the visual field (taken from Rayner et al. 2012).

2.4.2 Theoretical Advancements

The surge of interest in cognitive psychology in the 1960s and 1970s motivated scientists to build on basic relationships discovered between eye movements and visual stimuli, and consider the complex cognitive processes that they represent (Jacob and Karn 2003). The study of eye movements during reading tasks continued, and notable contributions from psychologists Marcel Just and Patricia Carpenter were made. Knowing that eye movements during reading consist of pauses (fixations) and jumps (saccades), Just and Carpenter (1976a, 1976b) investigated how these pauses related to cognitive processes. Their research led to the development of the eye-mind hypothesis, which was a major theoretical advancement within eye-tracking research. This assumption states that "…there is no appreciable lag between what is being fixated and what is being processed" (Just and Carpenter 1976b). Using this hypothesis, Just and Carpenter suggested other ways to utilize measures of eye fixations for studying cognitive processes, including: problem solving, spatial information processing, and real-world scene processing (Just and Carpenter 1976b).

While the eye-mind hypothesis has been supported extensively within the eye-tracking research community, its limitations are important to consider. One limitation is that laboratory research has shown attention to precede fixations slightly, such that attention shifts approximately $250\,ms$ before the eye moves (Deubel 2008). Although studies have not confirmed whether this lag also exists in more natural settings, we cannot expect the entire duration of a fixation to be representative of the information that is being acquired and processed (Holmqvist et al. 2011). Despite this observed short lag, though, scientists believe that eye movements and attention are still tightly coupled, and we can expect the eye to follow where attention is redirected to (Holmqvist et al. 2011). Another notable limitation is that information can be processed for some time after it has been fixated. Just and Carpenter (1976b) explained that the eye-mind hypothesis describes the processing of fixations that are "…at the top of the stack in active memory" (Just and Carpenter 1976b). This statement supports that we may have additional items in our active memory, although they may not be the focal of our attention.

2.4.3 Technology Advancements

Eye-tracking technology in the late 19th and early 20th centuries was very basic. While some types of equipment were invasive and required direct mechanical contact with the eye, others made use of motion picture photography techniques. For example, Judd et al. (1905) inserted a small white particle into the eye before taking a series of photographs of it. Geometrical illustrations using points of reference from worn spectacles and the white particle were drawn from photograms to determine eye movement. Methods using corneal reflections from light source and motion pictures were later developed (Mackworth and Mackworth 1958). While these methods were far less invasive, restraining movement was still important; participants were therefore asked to bite on a plastic mold and use a rigid head and cheek bone support to prevent movement (Fig. 2.7) (Mackworth and Mackworth 1958).

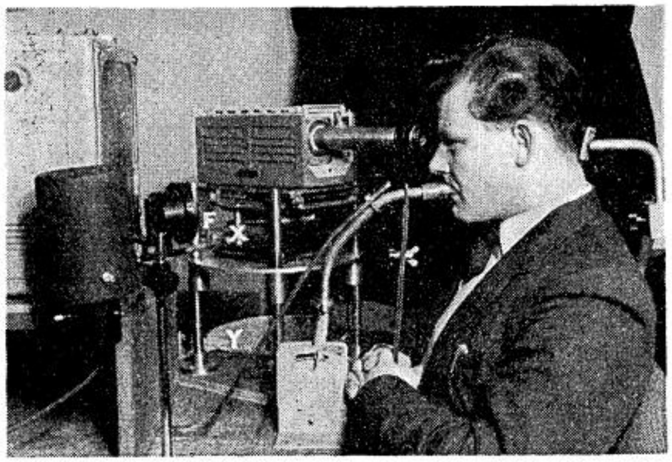


Figure 2.7. Eye tracking setup using corneal reflections, a camera, and movement restraint (taken from Mackworth and Mackworth 1958).

Modern day eye-tracking systems have improved substantially compared to the invasive and uncomfortable equipment used in the beginning stages of eye-tracking research. Today’s systems use pupil center corneal reflection methods. The use of infrared light to illuminate the eye is advantageous because it is not visible to the person whose eyes are being illuminated. Depending on the selected system, video-based cameras capture images of a person’s eyes and their corneal reflections at a sampling rate of $25 - 2000\ Hz$ (Bojko 2013). With the use of image-processing algorithms and a three dimensional model of the eye, the location of a person’s gaze can be determined. A sampling rate of $250\ Hz$ is sufficient for detection of very small eye movements (such as saccades) and for measuring the duration of eye movements within $\pm 2\ ms$ (Bojko 2013).

Diodes that emit infrared light can be positioned in the camera in two ways: either in line with the optical axis of the camera (bright-pupil eye tracking), or away from the optical axis of the camera (dark-pupil eye tracking) (Bojko 2013). Bright-pupil eye tracking causes the pupil to appear brighter than the iris because the reflected infrared light from the retina is in line with the camera. Conversely, dark-pupil eye tracking causes the infrared light to be reflected away from the camera, and thus the pupil appears darker than the surrounding iris. Though both approaches are designed to illuminate the eyes and create contrast between the iris and pupil, they perform differently depending on the ambient lightning conditions and physical characteristics of the eye. Bright-pupil eye tracking is best suited in darker environments when the pupil tends to be less dilated, and is more effective at tracking people with bright colored eyes (i.e., blue). Dark-pupil tracking, on the other hand, works well in most lightning conditions and is more effective at detecting the pupil in dark colored eyes (i.e., green and brown). However, one caveat to dark-pupil eye tracking is that surrounding dark features can disrupt pupil detection (such as eye lashes and the use of dark makeup).

Both head-mounted and remote-based eye-tracking systems are available for research use (Goldberg and Wichansky 2003). Head-mounted systems are worn, and while some devices are bulky and obstruct a person’s view, others have simpler designs that resemble a pair of glasses (Fig. 2.8a). While these systems allow for large head movements during data collection and provide a way for observing eye behavior in natural settings, they can be uncomfortable and do not allow participants to engage in tasks without being fully aware that their eyes are being observed. Remote-based systems are typically positioned beneath or within a computer monitor that display the task at hand

(Fig. 2.8b). Advancements in this technology now allow for some head movement, which makes this method less invasive because the use of chin rests to stabilize participants during observation is not necessary. An additional advantage to remote-based eye-tracking systems is that the observed visual scene stays within the same boundaries for the duration of an experiment; these set boundaries make data analysis much more straightforward. With head-mounted systems, content within a participant's gaze position is constantly changing, which can make data analysis challenging and time consuming.

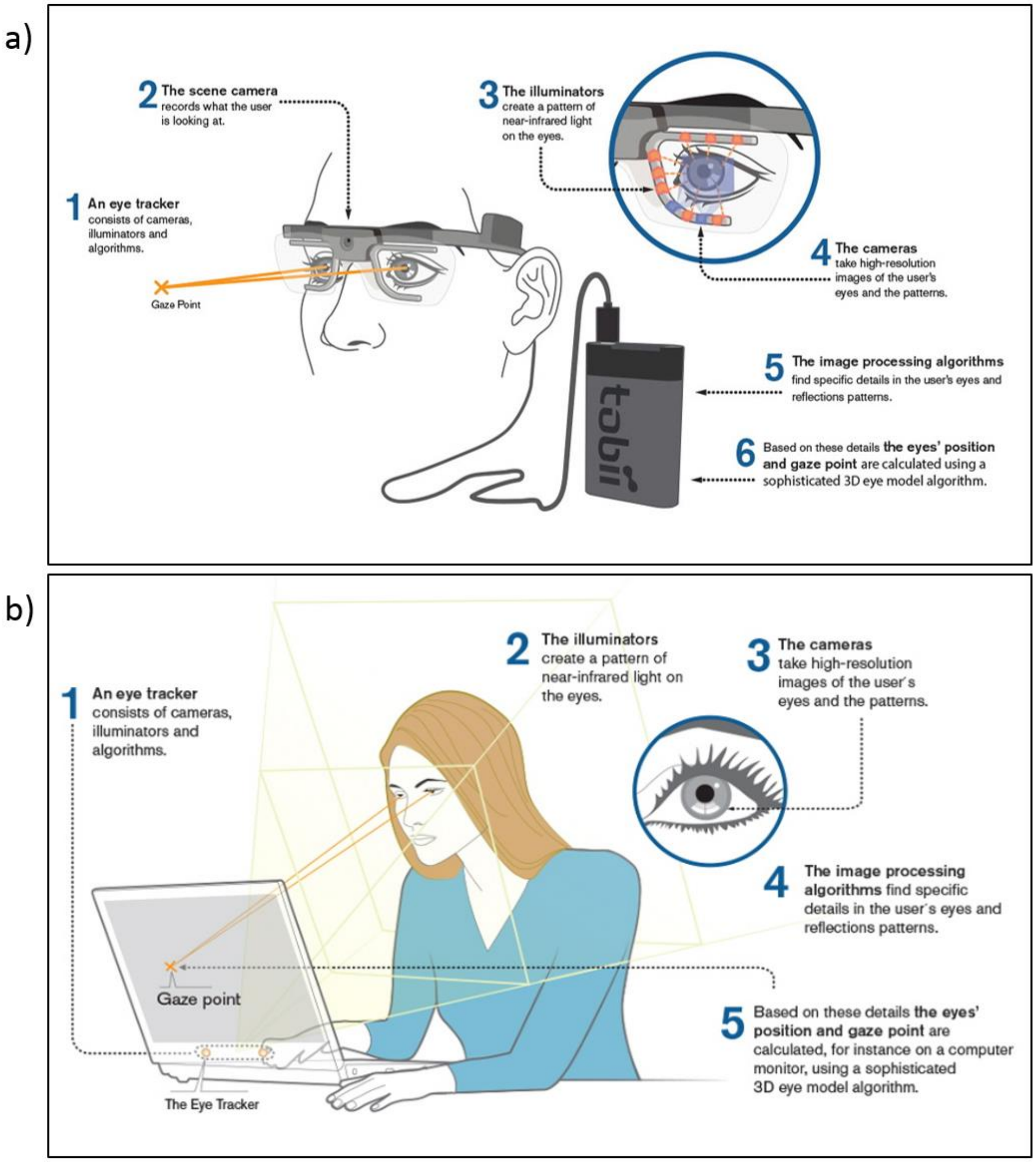

a)
2 The scene camera records what the user is looking at.
3 The illuminators create a pattern of near-infrared light on the eyes.
1 An eye tracker consists of cameras, illuminators and algorithms.
Gaze Point
4 The cameras take high-resolution images of the user's eyes and the patterns.
tobii
5 The image processing algorithms find specific details in the user's eyes and reflections patterns.
6 Based on these details the eyes' position and gaze point are calculated using a sophisticated 3D eye model algorithm.
b)
2 The illuminators create a pattern of near-infrared light on the eyes.
3 The cameras take high-resolution images of the user's eyes and the patterns.
1 An eye tracker consists of cameras, illuminators and algorithms.
4 The image processing algorithms find specific details in the user's eyes and reflections patterns.
Gaze point
5 Based on these details the eyes' position and gaze point are calculated, for instance on a computer monitor, using a sophisticated 3D eye model algorithm.
The Eye Tracker

Figure 2.8. Illustrations of eye gaze position determination using a) a head-mounted eye-tracking system and b) a remote-based eye-tracking system (Tobii 2017).

In the work presented in this dissertation, a remote video-based Tobii TX300 eye-tracking system was used with dark-pupil methods (Fig. 2.9). The sampling frequency was 300 Hz. It allowed for head movement, meaning that a chin rest or other form of restraint was not required. Seated at a distance of 65 cm from the eye-tracking system, participants could move 37 cm in width and 17 cm in height. Within these bounds, the eye tracker could detect at least one eye. The gaze accuracy, referring to the possible angular distance error from actual to observed point of gaze, is on average 0.4 ° (Tobii 2014). This accuracy corresponds to a 4.8 mm possible error in gaze location on the computer screen. The gaze precision of this system, referring to the spatial angular variation between gaze samples, is 0.07° (Tobii 2014).

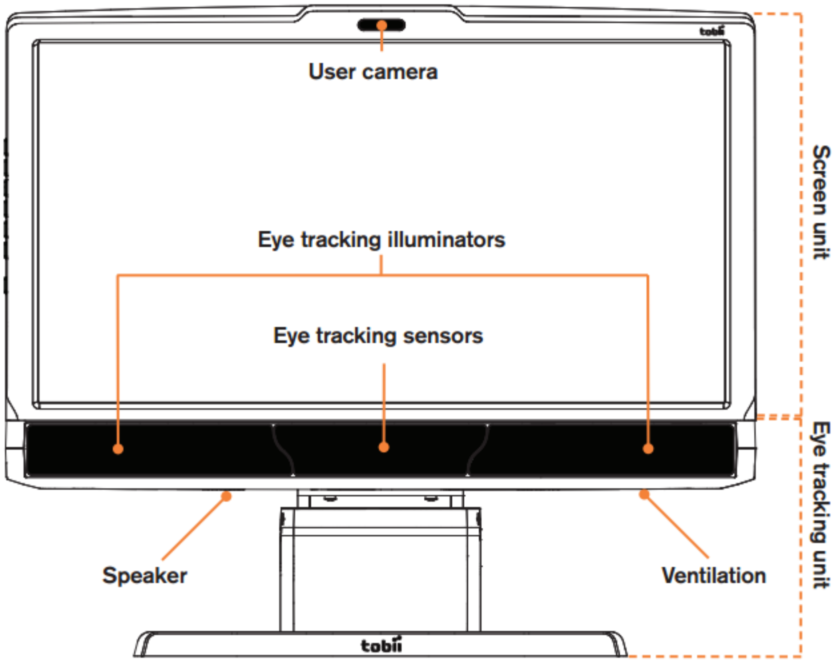


Figure 2.9. Front display of the Tobii TX300 system (Tobii 2014).

2.4.4 Making Sense of Eye-Movement Data

The technological and theoretical advancements related to eye-tracking research have led to tremendous growth in applications of this technology to a variety of research domains. In turn, this growth has resulted in the development of numerous analysis approaches to making sense of eye-movement data. The early discovery that information is acquired and processed during eye fixations continues to be central to data analysis today. The majority of results that are reported in research studies describe fixation activity, such as how often and for how long a person looks at a piece of information. Most eye-tracking system software provide algorithms that can identify different types of eye activity, including fixations and saccades, and these results can be presented in either a qualitative or quantitative manner.

Qualitative representations of eye activity can be visualized in a number of different ways (Bojko 2013). Heatmaps illustrate aggregate eye movements over a set time, and depict the spatial distribution of visual attention in a static image. Spatial distributions of both the frequency and the absolute and relative duration of fixations can be plotted. However, heatmaps do not provide temporal information about fixation behavior. Gaze plots, on the other hand, can depict a series of individual fixations with the linking saccades in a static image. Additionally, in these plots, the size of each dot indicates the duration of each fixation. Taken a step further, gaze plots can be turned into videos, and the order of fixations relative to the background content can be replayed.

While qualitative representations of eye movements give insight into how a person distributes their visual attention, quantitative representations are more useful for providing measures of that behavior. In an analysis of the measures reported in 21 usability studies, Jacob and Karn (2003) found that the overall number of fixations (count), gaze percent (proportion of time), and the overall mean fixation duration were used most frequently. Areas of interest describe different portions of a scene and semantic content usually determines the area borders. Interpretation of fixation measures depends on the task at hand, but typically, information that is fixated on more frequently is considered more important or noticeable to the viewer, and information that is fixated on for a longer duration is considered more engaging or more difficult to process (Poole and Ball 2006; Bojko 2013). Summaries of these measures can be presented in graphical forms, and inferential statistics can be used to test for statistical significant differences between treatment conditions.

Many of the quantitative measures provide bulk summaries of fixation behavior, but they do not capture how fixations evolve over time. The temporal aspects of eye movements are important because they represent underlying cognitive processes driving the ways in which attention is focused. Therefore, scientists have developed methods that consider and compare sequences of fixations. Noton and Stark (1971) first termed these sequences a scanpath, which more recently was given the physical definition of "the route of oculomotor events through space within a certain timespan" (Holmqvist et al. 2011). A variety of methods to compare different aspects of scanpaths have been developed, and each have their own advantages and limitations (Anderson et al. 2015). Some of these methods include the String Edit Distance (Levenshtein 1966), ScanMatch (Cristino et al. 2010), and MultiMatch (Jarodzka et al. 2010) algorithms. The latter of these may be considered the most comprehensive method given that it compares five characteristics of scanpaths (vector, length, direction, position, and duration), whereas other methods compare just one or two characteristics (Anderson et al. 2015). Scanpath comparison algorithms output results in terms of similarity scores; these scores can be used to determine how well-mapped the underlying cognitive processes represented by two scanpaths are.

2.4.5 Broadening the Research Applications

Eye-tracking research originates in the study of reading and the desire to understand language processing mechanisms. However, scientists recognized that eye movement observations could be useful for other types of tasks. Fitts et al. (1950) was the first to apply eye-tracking methods to a usability study in an aircraft landing approach task. In this study, a motion-picture camera was used to capture pilots' eye movements during flight, and these film records were analyzed to determine the importance of different instruments in the cockpit, how difficult it was to interpret the instrument readouts, and how well the instruments were arranged relative to one another based on the spatial order of pilots' fixations. Despite this demonstration of a successful eye-tracking application in a usability study, this type of research was slow to take off because of technological and practical challenges (Jacob and Karn 2003). Eye-tracking systems were not easy to use and the algorithms that identify different types of eye activity were not yet developed. Furthermore, the dynamic scenes captured in usability studies made for more labor-intensive analyses and difficulties in interpretation (Jacob and Karn 2003).

Advancements in technology and theoretical understanding of eye movements coincided with the dramatic increase in computer use in the workplace, and a new research avenue was identified: human-computer interaction. This new research opportunity meant that applications of eye tracking to usability studies was a natural next step that many scientists embraced (Duchowski 2002; Jacob and Karn 2003). One area in which eye tracking was first applied to usability studies was in assessing how pull-down menus on computer screens are used (e.g., Card 1984; Hendrickson 1989; Aaltonen et al 1998; Byrne et al. 1999). Observing the ways in which users interact with these menus and make choices is helpful for improving the design of graphical interfaces so that they are easier and more efficient to navigate. Using eye tracking to observe how people search web pages either freely or with a goal in mind has also helped inform design recommendations for the layout of websites (Benel et al. 1991; Ellis et al. 1998; Goldberg et al. 2002).

The use of eye tracking to test how people interact with computer displays was not constrained to the office environment. Following from Fitts et al.'s (1950) first application of eye tracking in a pilot usability study, other researchers adopted similar methods to tap into the cognitive processes of pilots. In a study examining the impacts of information complexity on combat pilots, Svensson et al. (1997) found that eye movements were sensitive to the amount of information available. Furthermore, Flemisch and Onken (2000) observed six military pilots' eyes while they completed a navigation task in a flight simulator with different types of information displays. The pilots' eye movement data were useful for evaluating their distributions of visual attention across the different displays and for considering the best way to provide technical support (Flemisch and Onken 2000). Tracking pilots' eyes in the cockpit has continued over the years, and has more recently provided insight into pilots' situational awareness during a malfunction in a simulated flight scenario (Van De Merwe et al. 2012), and for observing differences in expert and novice pilots' distributions of attention and related eye movements during flight (Sullivan et al. 2011; Yu et al. 2016).

Eye-tracking applications within the aviation domain have not been reserved to just the cockpit; eye tracking has also been an effective tool for studying air traffic controllers. For example, Hauland (2008) used eye tracking to determine air traffic controller students' situational awareness during a flight simulation based on how their attention was distributed across the display and how they acquired information. Differences in novice and experts' sequences of fixations was also of interest in a conflict detection task (Kang and Landry 2014). Additionally, Kang and Landry (2014) showed that after novices viewed how experts' eye movement sequences on the air traffic control display, their number of false alarms reduced. Use of these sequences was therefore considered effective for training purposes.

In addition to air traffic control, another research area that has recognized opportunities for applying eye-tracking methods is the medical field. Due to practical matters, many of these applications have focused on medical imaging studies, in which the visual search behavior of medical professionals is analyzed (Al-Moteri et al. 2017). Such observations have allowed researchers to determine whether incorrect diagnoses of abnormalities in medical images occur at the detection or decision stage (Mannging et al. 2014). Furthermore, a popular use of eye tracking within the medical community has been to examine differences in novices and more experienced radiographers' visual search behavior when tasked with detecting abnormalities in medical images (Wood et al. 2013; Giovinco et al. 2015; Bertram et al. 2016). For example, Wood et al. (2013) tracked the visual search behavior of radiologists tasked with detecting and diagnosing fractures in radiographs. In this study, the more experienced radiographers were found to fixate on fractures more quickly and spent more time fixating on the fracture area than less experienced radiographers (Wood et al. 2013). Though not as common, differences in novice and experienced surgeons' eye movements during simulated and live operations have also been investigated (Tien et al. 2010; Zheng et al. 2011; Khan et al. 2012; Tien et al. 2015). The knowledge developed from these types of research efforts in both the aviation and medical fields are important for the development of training material that will promote a safer, more efficient, and better performing workforce. Researchers within education have also more recently recognized the possibilities of using eye movement data from experts as a training tool for teaching and learning (Jarodzka et al. 2017).

Eye tracking has also been used in marketing research, which presents a less stressful and consequential decision environment than aviation or medicine. One example of a topic in marketing that eye tracking has been useful for understanding is banner blindness. This term refers to the tendency to ignore advertisements that pop up on web pages. Eye tracking has shown that the types of tasks that users engage in online as well as the location of advertisements on web pages modifies the extent to which banner blindness exists (Albert 2002; Hervet et al. 2011; Resnick and Albert 2013). Eye-tracking results have shown that people are less likely to observe advertisements when completing goal-oriented tasks or when the banner is positioned on the right hand side of the web page (Albert 2002). Additionally, eye-tracking studies within supermarkets have provided insight into the ways in which package design influence shoppers' visual attention and how shoppers choose products (Clement et al. 2013; Gidlöf et al. 2013).

2.4.6 Eye Tracking for Meteorology

With the majority of meteorology research focusing on physical aspects of the atmosphere, it is unsurprising to learn that eye tracking has been used on very few occasions in this field. The earliest reported application of eye tracking in meteorology was for learning more about how humans extract implicit and explicit information from complex visualizations. In this study, Trafton et al. (2002) presented meteorological visualizations that varied in terms of completeness to United States' Navy weather forecasters and tracked their eye movements as they answered basic quantitative and qualitative questions. The findings showed how forecasters use spatial representations to interpolate between lines (i.e., isobars) and that even experienced users of these visualizations refer to legends often (Trafton et al. 2002). More recently, Sherman-Morris et al. (2015) used eye tracking to investigate how altering meteorological visualizations impacts users' graph comprehension. Eye movement data showed that both the shading color and the units used in the graph's legend affected participants' abilities to interpret hurricane storm surge graphical information correctly (Sherman-Morris et al. 2015). Furthermore, Drost et al. (2015) applied eye tracking in a recent broadcast meteorology research project to learn about how a weathercaster's gesturing can affect viewers' attention during a televised weather forecast. Viewers' eye movement data showed that the weathercasters' gesturing affected where their attention was directed but not their retention of information.

These listed studies in meteorology show that eye-tracking methods have been useful for learning about graph comprehension and assessing how well presentations of meteorological information are conveyed to users. However, studies have not utilized eye-tracking technology to learn more about NWS forecasters and their warning decision processes. The wide ranging applications of eye tracking across the medical, aviation, and marketing worlds, as well as the rare but successful uses within meteorology, suggest that eye tracking has strong potential for use in this field of research. Research questions that eye tracking has helped answer in medical imaging studies are certainly analogous to those we may ask in meteorology. For example, visual search tasks requiring radiographers to detect fractures in x-rays is analogous to requiring forecasters to detect radar signatures indicative of a specific weather threat. Drawing on these effective uses of eye tracking, research methods can be adapted and applied within the PARISE setting to better observe and learn about forecasters' warning decision processes as they interrogate, process, and act on radar data.

Chapter 3

Forecaster Performance and Workload: Does Radar Update Time Matter?

Taken in full from: Wilson, K. A., P. L. Heinselman, C. M. Kuster, and D. M. Kingfield, 2017: Forecaster performance and workload: Does radar update time matter? *Wea. Forecasting*, **32**, 253–274.

Abstract

Impacts of radar update time on forecasters' warning decision processes were analyzed in the 2015 PARISE. Thirty NWS forecasters worked nine archived PAR cases in simulated real time. These cases presented nonsevere, severe hail and/or wind, and tornadic events. Forecasters worked each type of event with approximately 5-min (quarter-speed), 2-min (half-speed), and 1-min (full-speed) PAR updates. Warning performance was analyzed with respect to lead time and verification. Combining all cases, forecasters' median warning lead times when using full-, half-, and quarter-speed PAR updates were 17, 14.5, and 13.6 min, respectively. The use of faster PAR updates also resulted in higher Probability of Detection and lower False Alarm Ratio scores. Radar update speed did not impact warning duration or size.

Analysis of forecaster performance on a case-by-case basis showed that the impact of PAR update speed varied depending on the situation. This impact was most noticeable during the tornadic cases, where radar update speed positively impacted tornado warning lead time during two supercell events, but not for a short-lived tornado

occurring within a bowing line segment. Forecasters' improved ability to correctly discriminate the severe weather threat during a nontornadic supercell event with faster PAR updates was also demonstrated. Forecasters provided subjective assessments of their cognitive workload in all nine cases. On average, forecasters were not cognitively overloaded, but some participants did experience higher levels of cognitive workload at times. A qualitative explanation of these particular instances is provided.

3.1 Introduction

During convective warning operations, NWS forecasters rely primarily on weather radar to monitor storms and make warning decisions. The WSR-88D network currently provides forecasters with volumetric updates every 4–6 min. However, given that PAR may likely become the next generation of weather radar, this technology is being tested and considered for weather applications (Forsyth et al. 2005; Zrnić et al. 2007). Located in Norman, Oklahoma, the National Weather Radar Testbed PAR (hereafter PAR) demonstrates how electronic beam steering can be used to adaptively scan the atmosphere and collect rapid-update (~1 min) volume scans of a 90° azimuthal sector (Heinselman and Torres 2011).

In a continued effort to improve the timeliness and accuracy of warnings, it is vital that the potential impacts of higher-temporal resolution radar data on NWS forecasters' warning decision processes are understood. Since 2010, PARISE has been addressing a variety of research questions to examine this issue (Heinselman et al. 2012, 2015; Bowden et al. 2015; Bowden and Heinselman 2016). Applications of behavioral science methods (e.g., cognitive task analysis) have resulted in a better understanding of forecasters' thought processes as they interrogate radar data and make warning decisions. This analysis has provided important insight into aspects of forecasters' performance, such as lead time and verification, which itself has been a consistent focus throughout PARISE. Impacts of 1-min PAR updates on forecasters' performance during a variety of scenarios was assessed in the 2010, 2012, and 2013 PARISE.

The 2010 PARISE focused on a known challenge within the NWS: being able to provide warning lead time on weak, short-lived tornadoes. Comparing forecasters' decisions when using 43-s versus 4.5-min volumetric PAR updates, this experiment found that participants using faster updates achieved longer tornado warning lead times (Heinselman et al. 2012). However, forecasters using these faster updates also had a higher False Alarm Ratio (FAR). Due to the small sample size in the first experiment and the concern that faster PAR updates could lead to a higher number of false alarms, the experimental design was modified in the 2012 PARISE and the number of cases that participants worked was increased (Heinselman et al. 2015). This time, forecasters worked a total of four events (two tornadic and two nontornadic) independently, each with 1-min updates. The participants achieved a median tornado warning lead time of 20 min, which exceeded the EF0/EF1 tornado warning lead time of the participants' respective forecast offices (7 min) and NWS regions (8 min; Heinselman et al. 2015). All but one forecaster also achieved a probability of false alarm score <0.5, indicating that warning accuracy was better than chance during this experiment (Heinselman et al. 2015).

Although the 2010 and 2012 PARISE results demonstrated positive impacts of higher-temporal resolution radar data on forecasters' warning decisions during weak tornado events, a question that remained was whether the same benefits would be observed during events that only produced severe hail and/or wind. The 2013 PARISE aimed to answer this question using a two-independent group design, such that half of the participants were each assigned to a control group (5-min updates) while the other half was assigned to an experimental group (1-min updates). Performance of the experimental group during these cases was superior to that of the control group, as demonstrated by their statistically significant longer median warning lead time (21.5 min) compared to the control group's (17.3 min), and their more accurate warning decisions (Bowden et al. 2015).

Previous PARISE studies have contributed substantially to our understanding of the potential impacts of higher-temporal resolution radar data on forecasters' warning decision processes. However, there have been some key limitations preventing the generalizability of our findings about forecasters' performance. The most notable limitation is the sample size; in each PARISE, only twelve forecasters were recruited for participation and only 1–4 cases were worked. In each experiment, these cases focused on a specific weather threat (i.e., weak tornado or severe hail/wind), and as a result they did not provide the variety of weather events typical in a forecast office. Furthermore, while impacts of 1-min and 5-min PAR updates have been explored, we have not assessed how forecasters would perform with 2-min PAR updates. Finally, forecasters' cognitive burden resulting from a greater influx of data was not examined in these previous experiments, and therefore the effects of rapidly-updating PAR data on forecasters' cognitive workload was still unknown.

The 2015 PARISE was therefore designed to address these limitations, while continuing to deepen understanding of forecasters' warning decision processes and target new research questions. Based on findings from previous experiments, we expected forecasters with faster PAR updates to perform better, most notably with respect to warning lead time. We also expected forecasters with faster PAR updates to discriminate between weather threats more successfully. Given that forecaster cognitive workload had not been studied in detail in the literature, we were hopeful that our assessment would provide new insight into forecasters' mental efforts during warning operations. Our expectation was that faster PAR updates would lead to increased cognitive workload, especially during more demanding weather scenarios. In this paper, we provide an overview of the experimental design and methods applied in the 2015 PARISE. We focus our analysis on how forecasters' performance, warning characteristics, and perceived cognitive workload relate to the temporal resolution of radar data and the type of weather threat presented in each case. Finally, we bring together findings from this most recent study and from previous studies to give an overall assessment of what higher-temporal resolution radar data will mean for NWS forecasters during warning operations.

3.2 Methodology

The 2015 PARISE took place over six weeks during August and September 2015. Each week, five NWS forecasters visited the NOAA Hazardous Weather Testbed in Norman, Oklahoma, and completed three experimental components of this study. These components were the traditional experiment, eye-tracking experiment, and focus group. The traditional experiment built directly on earlier PARISE studies, aiming to improve the generalizability of PARISE findings through increased sample size of participants and cases worked. Additionally, the traditional experiment explored the concept of cognitive workload for the first time in PARISE. This paper discusses findings from the traditional experiment only.

3.2.1 Recruitment

Thirty NWS forecasters were recruited for the 2015 PARISE. Since forecasters would be working archived weather events from central Oklahoma, those most likely to have encountered similar storm types during their own warning operations were targeted. The 30 participating forecasters represented 25 NWS WFOs located across eleven states in the Great Plains (Fig. 3.1). Of these forecasters, 5 were female and 25 were male, and experience ranged from 1–27 years (mean=12 years, SD=7 years). Prior to participating in this study, all forecasters completed a multiple choice survey that was comprised of 48 questions drawn from forecaster training material designed by the NOAA Warning Decision Training Division. This survey queried forecasters' knowledge of severe weather definitions and their understanding of conceptual models and weather radar. The purpose of this survey was to obtain a simplistic assessment of forecasters' general knowledge of severe weather warning operations, which when represented as survey scores, could be used as a measure for comparison. The survey scores ranged from 28–41 out of a possible 49 points (mean=36, SD=3).

3.2.2 Experimental Design

A goal of the 2015 PARISE was for all forecasters to work a variety of weather events and to be exposed to a variety of temporal resolutions of PAR data. In comparison, each previous PARISE study was confined to a single type of weather (i.e., weak tornado events only or severe hail and wind events only), and forecasters were assigned to work with only 1-min or 5-min PAR volumetric updates (Heinselman et al. 2012, 2015; Bowden et al. 2015). This current study continued the assessment of forecaster use of 1-min and 5-min PAR volumetric updates, but based on forecasters' suggestions during the 2013 PARISE, also tested forecasters use of 2-min PAR volumetric updates (Bowden and Heinselman 2016).

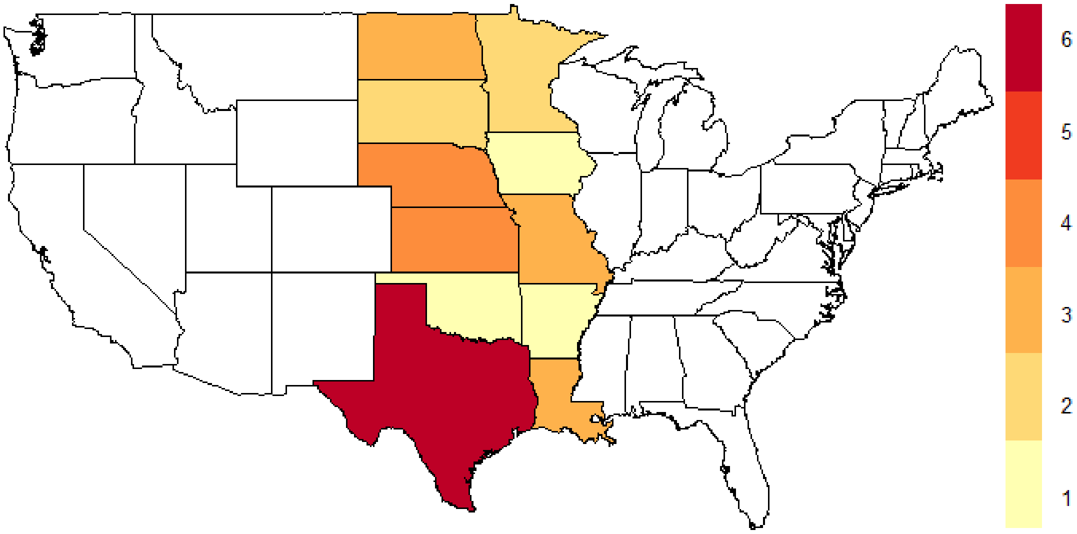


Figure 3.1 Forecasters were recruited from the Great Plains region of the United States. The color bar indicates the number of forecasters participating from each of the eleven states.

To examine forecaster use of these three temporal resolutions (full-speed [~1-min], half-speed [~2 min], and quarter-speed [~5 min]) for different types of weather events, nine archived PAR cases were selected (see section 3.3). The chosen experimental design required random assignment of forecasters to three separate groups, and each group was comprised of ten forecasters. Group assignment determined the temporal resolution of PAR data that would be used for each case, and all participants were exposed to the full-, half-, and quarter-speed PAR updates for each of the three case types.

3.2.3 Methods

3.2.3.1 Working Events

The majority of forecasters' participation time was spent on the traditional experiment. Forecasters worked on two to three cases per day, and the nine cases were completed in random order to avoid order effect. Forecasters were provided with their own AWIPS-2 workstations and worked each case independently. They did not discuss details of the weather events with other participants until the end of the week. First, a practice case was completed to train forecasters on how to setup their cases and to ensure that they were comfortable loading and interrogating PAR data in AWIPS-2. During this initial case, forecasters practiced issuing warnings using the Warning Generation (WarnGen) software, practiced receiving storm reports, and personalized settings in AWIPS-2.

Similar to previous PARISE studies, prior to working each case forecasters viewed a pre-briefing video that described the environmental conditions associated with the upcoming case. Mesoscale analysis, sounding information, and satellite and radar data were provided, and forecasters used this information to form and document their expectations for how the event might unfold. When working the case, forecasters were able to view reflectivity, velocity, and spectrum width products in simulated real time. Importantly, forecasters were asked to work the event in their normal forecasting style, and to interrogate the radar data and issue special weather statements, warnings (severe thunderstorm and tornado), and severe weather statements that they deemed necessary. All issued products were recorded in a database for performance analysis.

3.2.3.2 Workload Ratings

With an increase in data availability, the impact of higher-temporal resolution radar data on forecasters' workload is of interest. Workload is defined as the level of attention resources required to meet the performance criteria, and is affected by task demands and past experience (Young and Stanton 2005). Widely used workload assessment methods are the NASA-TLX (Hart and Staveland 1998) and SWAT (Reid et al. 1981); however, both methods evaluate workload based on sub-classifications such as time demand, effort demand, and stress demand, which can be time consuming and obtrusive when workload needs to be evaluated many times during a prolonged task. Furthermore, given that forecasters' work demand is predominantly cognitive, many of these sub-classifications are difficult for forecasters to relate to. Thus, a faster, less obtrusive, and more suitable method was chosen. This method was the Instantaneous Self-Assessment (ISA) (Kirwan et al., 1997), which is based on a unidimensional scale and has five qualitative ratings of mental effort, including: under-utilized (1), relaxed (2), comfortable (3), high (4), and excessive (5) (Miller 2001). Each level of mental effort was provided with a corresponding description. The ratings can also be thought of in terms of how much spare mental capacity one has (Table 3.1) (Kirwan et al. 1997). To capture variations in forecasters' mental workload during events, ISA ratings were collected during a video-cued retrospective recall at 5-min intervals. Along with each rating, forecasters provided reasoning for their chosen mental workload level.

Table 3.1. The Instantaneous Self-Assessment (ISA) tool adapted from Kirwan et al. (1997).

Level	**Workload**	**Spare Capacity**	**Description**
1	**Under-utilized**	**Very much**	Nothing to do. Rather boring.
2	**Relaxed**	**Ample**	More time than necessary to complete tasks. Time passes slowly.
3	**Comfortable**	**Some**	The controller has enough work to keep him/her stimulated. All tasks are under control.
4	**High**	**Very little**	Certain non-essential tasks are postponed. Could not work at this level very long. Controller is working 'at the limit'. Time passes quickly.
5	**Excessive**	**None**	Some tasks are not completed. The controller is overloaded and does not feel in control.

3.3 Radar Data

For the 2015 PARISE, the nine cases selected from archived PAR data maximized the variety in storm types, hazard types (e.g., severe hail), and distance from the radar. Each case also met temporal continuity (i.e., no data gaps) and duration criteria, which allowed forecasters ample time to demonstrate their warning decision process in each case. Following these criteria, we selected three null cases, three severe hail and wind cases, and three tornado cases based on storm reports provided by the National Centers for Environmental Information Storm Data publication (NCEI 2016).

Of the three null cases, two (Alpha and Epsilon) were multicell thunderstorms that produced no severe weather reports (Fig. 3.2a, b; Table 3.2). The third case (Theta) was considered null with respect to tornadoes. It contained two nontornadic supercells, but the supercell located about 75 km from the radar produced severe hail (Fig. 3.2c). In all three severe hail and wind events (Delta, Gamma and Beta), a multicell thunderstorm produced severe weather. In Delta, a storm produced both severe hail and wind, while storms in Gamma produced severe hail only and storms in Beta produced severe wind only (Table 3.2; Fig. 3.3). Of the three tornadic cases, Zeta contained a classic supercell that produced two tornadoes (one rated EF1 and the other rated EF2), Iota contained a supercell cluster that produced a tornado rated EF0, and Eta contained a tornadic squall line that produced a tornado rated EF1 (Fig. 3.4). The supercells in Zeta and Iota also produced severe hail and wind (Table 3.).

In all but one of the cases (Alpha), PAR operators collected data using a modified volume coverage pattern 12 (Brown et al. 2005) that included five additional elevation angles above 19.5° (up to 52.9°). For Alpha, a unique volume coverage pattern with 22 elevation angles between 0.51° and 52.94° was used. An adaptive scanning algorithm called ADAPTS (Heinselman and Torres 2011) was also used in all but three cases (Beta, Iota, and Alpha), which resulted in volumetric update times that varied throughout the cases (Table 3.2).

Table 3.2. Information about selected cases. Three case types are null, severe hail and/or wind, and tornadic.

Case	Time and date	Volume update time (s)	Storm type	Storm report
Null cases				
Epsilon	2059–2139 UTC 5 Jun 2012	55–60	Multicell	None
Alpha	0724–0756 UTC 14 May 2010	65	Multicell	None
Theta[a]	1957–2033 UTC 30 May 2013	67–78	Supercell	1.75-in. hail at 2013 UTC 1.0-in. hail at 2015 UTC 1.0-in. hail at 2015 UTC 1.0-in. hail at 2030 UTC
Severe hail and/or wind cases				
Beta	2120–2200 UTC 12 Aug 2011	72	Multicell	EG 61-kt wind at 2200 UTC
Gamma	2330–2359 UTC 22 Oct 2011	57–62	Multicell	1.50-in. hail at 2358 UTC
Delta	2222–2301 UTC 4 May 2012	32–40	Multicell	1.0-in. hail at 2242 UTC 1.25-in. hail at 2255 UTC MG 59-kt wind
Tornado cases				
Eta	2130–2200 UTC 29 May 2013	61–76	Squall line	EF1 tornado at 2200 UTC
Iota	2209–2301 UTC 30 May 2013	71	Supercell cluster	1.75-in. hail at 2212 UTC 4.25-in. hail at 2227 UTC EG 61-kt wind at 2230 UTC EF0 tornado at 2258–2259 UTC
Zeta	2050–2154 UTC 19 May 2013	64–76	Supercell	1.50-in. hail at 2101 UTC 1.25-in. hail at 2115 UTC EG 52-kt wind at 2115 UTC 1.0-in. hail at 2117 UTC 1.50-in. hail at 2118 UTC EF1 tornado at 2122–2130 UTC EF1 tornado at 2133–2134 UTC 2.60-in. hail at 2137 UTC EF2 tornado at 2141–2154 UTC

[a] Theta is a nontornadic hail-producing supercell (null tornado case).

Null

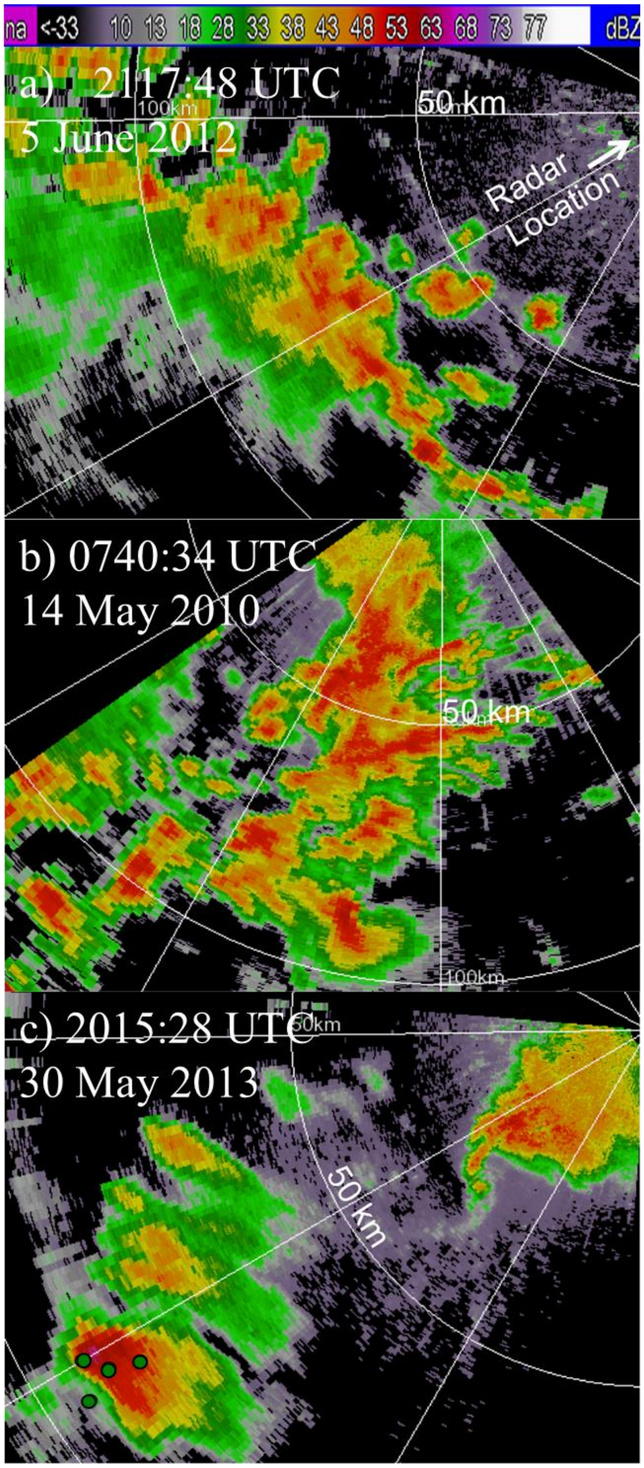


Figure 3.2. PAR 0.5° reflectivity for a) Epsilon, b) Alpha, and c) Theta (null tornado case). Green dots in c) are severe hail reports. Reflectivity (dBZ) color bar located at the top. White rings are displayed in 50-km increments.

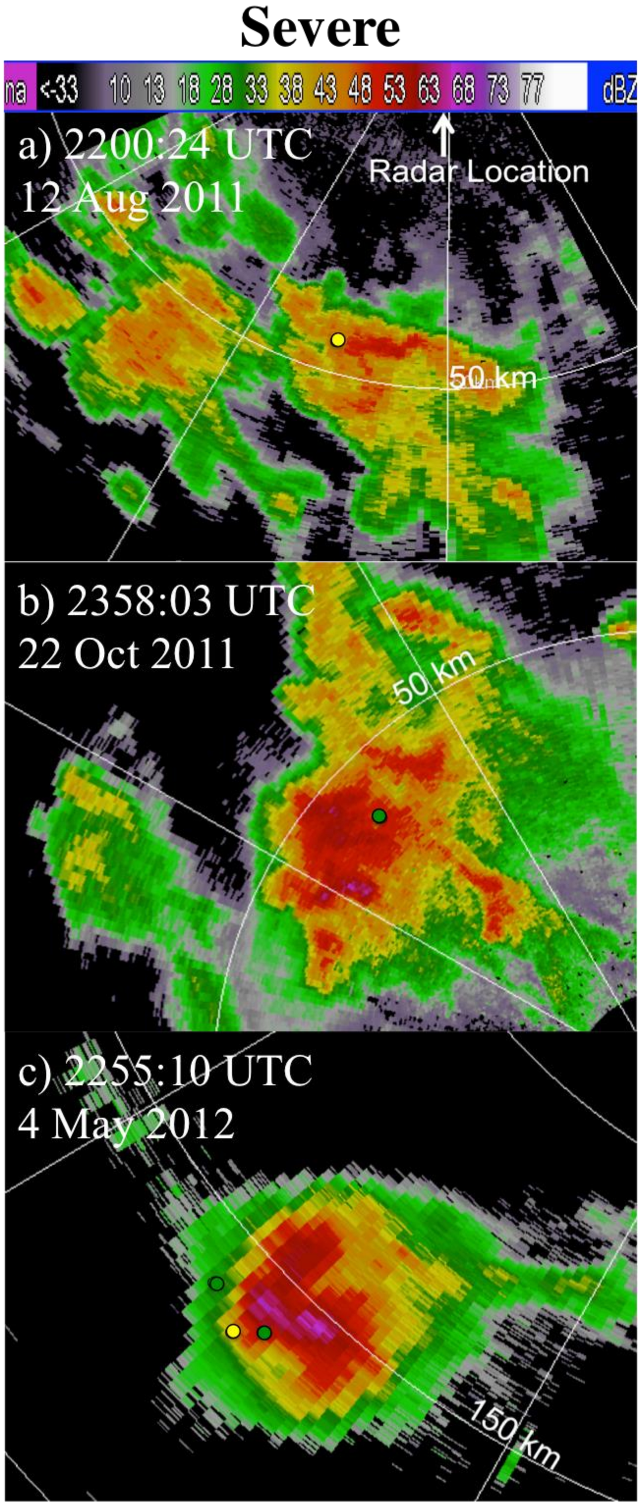


Figure 3.3. PAR 0.5° reflectivity for a) Beta, b) Gamma, and c) Delta. Green dots are severe hail reports and yellow dots are severe wind reports. Reflectivity (dBZ) color bar located at the top. White rings are displayed in 50-km increments.

Tornado

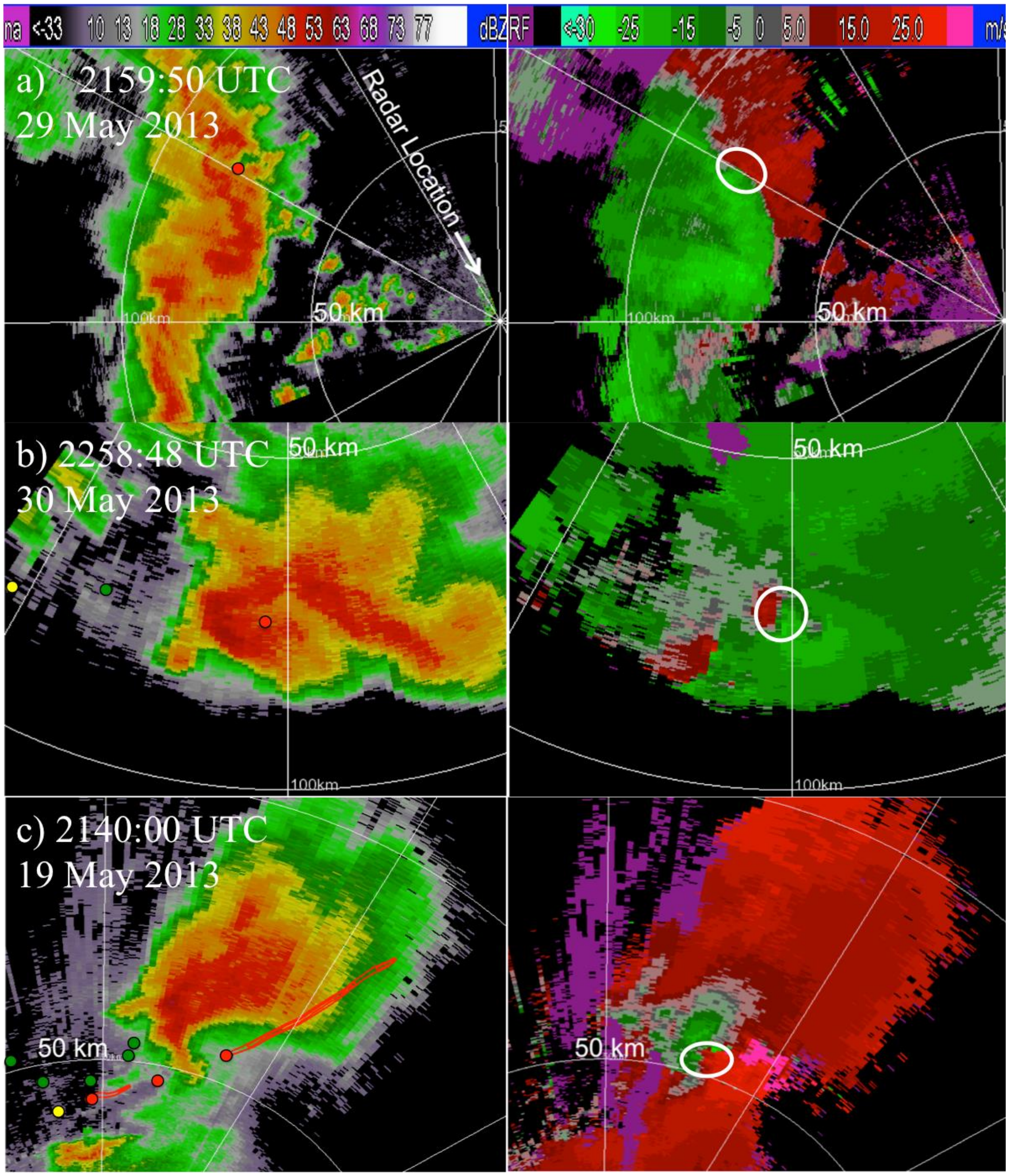


Figure 3.4. PAR 0.5° reflectivity (left) and velocity (right) for a) Eta, b) Iota, and c) Zeta. Green dots are severe hail reports, and yellow dots are severe wind reports. Red dots are tornado reports (i.e., starting point of tornado path), while red lines in c) are tornado paths associated with longer-lived tornadoes. Thick white circles show location of couplet of interest. Reflectivity (dBZ) and velocity (m s-1) color bars located at the top. White rings are displayed in 50-km increments.

3.4 Storm-Based Warning Verification

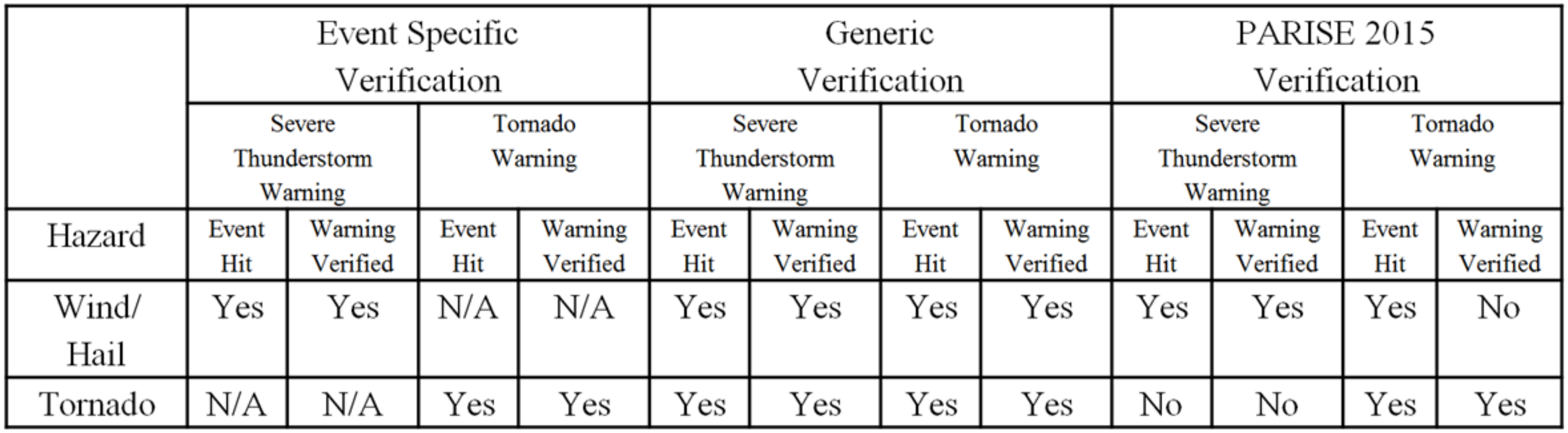
Recent PARISE experiments have focused on hazard-specific, storm-based warning verification of either tornadoes (Heinselman et al. 2012, 2015) or severe hail and winds (Bowden et al. 2015). In the 2015 PARISE, all three hazard types occurred in several of the simulation scenarios, requiring a verification framework for both severe thunderstorm and tornado warnings. As part of NWS Instruction 10-1601 (NWS 2015), two methods are used to verify these convective warnings: event specific and generic (Table 3.3). In the event-specific verification system, severe thunderstorm warnings are verified only by convective wind or hail events, and tornado warnings are verified only by tornado events. Because these matching hazard-to-warning combinations are only used to calculate hits and lead times, forecasters are neither rewarded nor penalized when an unmatched hazard-to-warning combination occurs. In the generic verification system, any convective hazard occurring in any warning type verifies the warning and allows for a lead time to be calculated for the hazard. Therefore, the generic verification system results in the possibility that a severe hail or wind event can verify a tornado warning and a tornado event can verify a severe thunderstorm warning.

For the above reasons, we decided to develop a hybrid verification system that adds certain components of the generic verification system to the event-specific verification system (Table 3.3). In this hybrid system, convective wind or hail events occurring within a tornado warning have their lead times calculated and count as a hit, but do not verify the warning. Wind or hail events occurring within a severe thunderstorm warning verify the warning and have event lead times tabulated, as they normally would. Tornado events occurring within severe thunderstorm warnings count as misses and do not verify the warning, with the opposite results occurring within a tornado warning. Our system allows for all events and warnings to be scored for each simulation but is stricter regarding tornado warning issuance and verification. In conjunction with the proposed hybrid verification system, we used the guidance within NWS directive 10-1601 (NWS 2015) to calculate the Probability of Detection (POD), FAR, and lead times for all warnings and hazards.

3.5 Performance

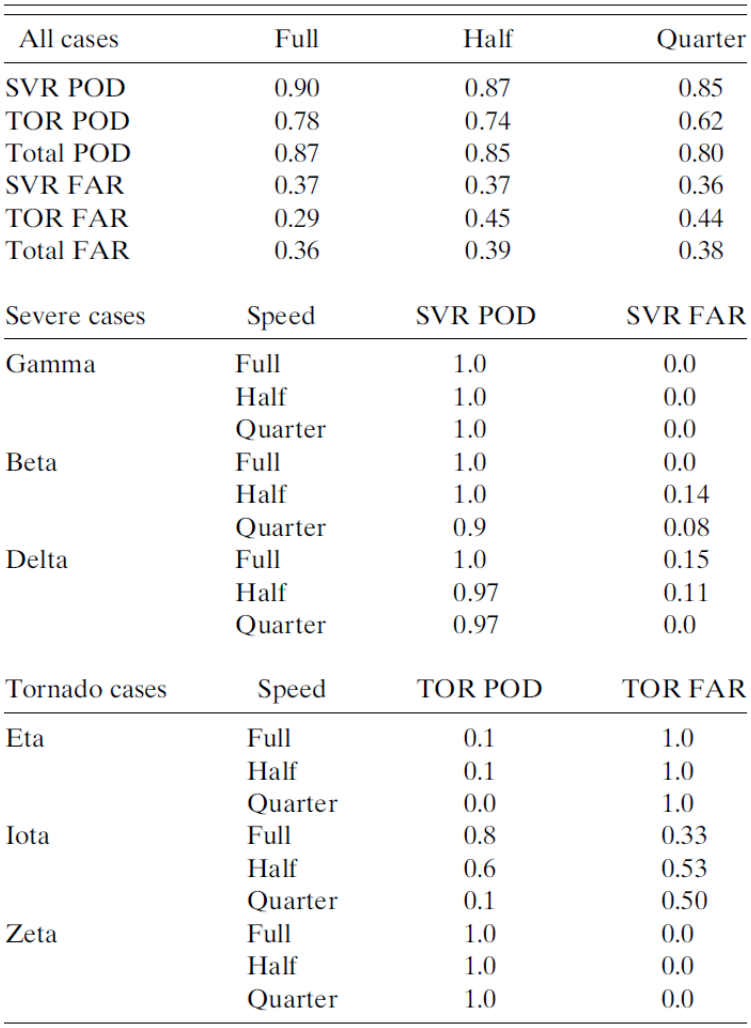
The expectation that overall median warning lead time would increase as update speed increased (became faster) was realized in this study. The use of full-, half-, and quarter-speed PAR data resulted in overall median warning lead times of 17, 14.5, and 13.6 min, respectively. Despite some difference in the median warning lead times, application of the Kruskal-Wallis test (Kruskal and Wallis 1952) showed no statistically significant differences between the three groups (*p*-value = 0.1683). This non-parametric test was chosen because the collected data did not meet normality assumptions. Overall POD and FAR scores were similar, with slight improvements as updates became more rapid (Table 3.4). Broken down by event type, the greatest differences are found for tornado warning POD and FAR scores (Table. 3.4). The full-, half-, and quarter-speed POD (FAR) scores were 0.78 (0.29), 0.74 (0.45), and 0.62 (0.44), respectively.

Table 3.3. The NWS event specific and generic verification systems used by the NWS and the hybrid verification system employed for PARISE.



	Event Specific Verification				Generic Verification				PARISE 2015 Verification			
	Severe Thunderstorm Warning		Tornado Warning		Severe Thunderstorm Warning		Tornado Warning		Severe Thunderstorm Warning		Tornado Warning	
Hazard	Event Hit	Warning Verified	Event Hit	Warning Verified	Event Hit	Warning Verified	Event Hit	Warning Verified	Event Hit	Warning Verified	Event Hit	Warning Verified
Wind/ Hail	Yes	Yes	N/A	N/A	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Tornado	N/A	N/A	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes

Table 3.4. The POD and FAR scores across all severe (SVR) and all tornado (TOR) warnings by update speed for all cases; severe cases Gamma, Beta, and Delta; and tornado cases Eta, Iota, and Zeta.



All cases	Full	Half	Quarter
SVR POD	0.90	0.87	0.85
TOR POD	0.78	0.74	0.62
Total POD	0.87	0.85	0.80
SVR FAR	0.37	0.37	0.36
TOR FAR	0.29	0.45	0.44
Total FAR	0.36	0.39	0.38
Severe cases	**Speed**	**SVR POD**	**SVR FAR**
Gamma	Full	1.0	0.0
	Half	1.0	0.0
	Quarter	1.0	0.0
Beta	Full	1.0	0.0
	Half	1.0	0.14
	Quarter	0.9	0.08
Delta	Full	1.0	0.15
	Half	0.97	0.11
	Quarter	0.97	0.0
Tornado cases	**Speed**	**TOR POD**	**TOR FAR**
Eta	Full	0.1	1.0
	Half	0.1	1.0
	Quarter	0.0	1.0
Iota	Full	0.8	0.33
	Half	0.6	0.53
	Quarter	0.1	0.50
Zeta	Full	1.0	0.0
	Half	1.0	0.0
	Quarter	1.0	0.0

These big-picture findings indicate that of the three update times used, full-speed data was most beneficial to forecasters' ability to issue more accurate warnings with longer lead times. However, of interest is how representative these findings are for each case worked. While examining this question, we found that the results were sensitive to the situation presented. For example, the temporal resolution used during Gamma and Eta had little impact on warning lead times, whereas differences were found in the other severe and tornado cases. Furthermore, in cases containing multiple reports, such as Delta and Zeta, we found the use of faster updates particularly improved warning lead times for the first report of the event. These longer initial warning lead times are an encouraging result, as warnings verified for the first report of the day tend to be the most challenging (e.g., Andra et al. 2002; Brotzge and Erickson 2009).

Given these situational dependencies, we expected that overall median warning lead times computed using only first reports from each case, and excluding Gamma and Eta, would show more improvement in warning lead time when using faster updates. Applying these criteria, the median lead times for full-, half- and quarter-speed were 14.5, 10.5, and 5.5 min, respectively (N=120 for each update-speed group). In this case, the application of the Kruskal-Wallis test did indicate statistically significant differences between the three groups (*p*-value = 0.0013). A post-hoc Wilcoxon-Mann-Whitney rank-sum test (e.g., Wilks 2006) indicates between which groups these statistically significant differences occurred (*p*-value<0.0170). Again, this non-parametric test was chosen because the data collected did not meet normality assumptions. Comparing the three groups, the full-speed group's median lead time distribution for this subset of the data was most different to that of the quarter-speed group's (*p*-value=0.0003), and provided additional confidence that the use of full-speed data did extend warning lead times compared to the use of quarter-speed data, in these cases. Further examination of first reports by case type revealed that the statistical significance found above was more so due to differences in tornado warning lead times between the three groups (Kruskal-Wallis *p*-value = 0.0380), rather than difference in severe thunderstorm warning lead times (Kruskal-Wallis *p*-value = 0.1162). The remainder of this section discusses the performance results by case type.

3.5.1 Performance: Severe Cases

The overall severe median warning lead times for the full-, half-, and quarter-speed cases were very similar: 21, 22.5, and 20 min, respectively (N=150 per group). As noted earlier, the most similar severe warning lead times occurred during Gamma, the hail-only case (Fig. 3.5a). Hence this case contributed to the overall similarity in median severe warning lead times found. To aid qualitative comparison between groups, in each case the median severe warning lead time for the full distribution (N=30) was computed. For Gamma, the full distribution lies near the 24.5-min median severe warning lead time. All groups achieved a perfect severe POD and FAR score (Table 3.4).

The most dissimilar severe warning lead times between the full-speed group and the quarter-speed group occurred during Beta, the wind-only event (Fig. 3.5b). Therein, both full- and half-speed groups achieved severe-warning lead times located mostly near or above the overall 18-min median lead time (N=30; Fig. 3.5b). In contrast, more than half of the quarter-speed group achieved severe-warning lead times at least 6-min under the 18-min median. The median severe warning lead times for full-, half-, and quarter-speed groups were 19.5, 18.0, and 10.5 min, respectively. The quarter-speed group's POD score was slightly lower and FAR score slightly higher compared to the full-speed group (Table 3.4). In this wind-only case, the use of half- and full-speed data was overall more advantageous to forecasters' ability to issue warnings with longer lead times than the use of quarter-speed data.

Unlike the other two cases, Delta contained both severe hail and wind reports. Because multiple storm reports were received as forecasters worked the case, warning lead times associated with the first report provided the clearest measure of the impact of temporal resolution on the warning decision process. As in Beta, groups using full- and half-speed data tended to issue warnings earlier (medians: 10 and 11 min, respectively) than the quarter-speed group (median: 6.5 min) (Fig. 3.5c). However, overall, the half-speed group outperformed the full-speed group, as the former produced the highest number of initial, second, and third severe warning lead times above the overall median warning lead times (10.5, 21.5, and 24.5 min, respectively) (Fig. 3.5c). One outlier was P29 of the half-speed group, who missed the first hail event; P9 of the quarter-speed group also missed the first event. The use of higher-temporal resolution data also resulted in slightly higher FARs compared to forecasters using quarter-speed data (Table 3.4).

3.5.2 Performance: Tornadic Cases

The overall median tornado warning lead times for the full-, half-, and quarter-speed cases were 12.7, 8, and 9 min, respectively (N=150 per group). Like the severe cases, performance for tornado cases was determined by the situation presented to forecasters. The most challenging tornado case for all groups was Eta, in which a short-lived EF1-rated tornado was produced on the north end of a bowing line segment approximately 75 km northwest of the PAR (Fig. 3.4a; Table 3.2). In this case, only 5 of 30 forecasters decided to issue tornado warnings prior to tornado occurrence: three were in the full-speed group (P11, P14, P15), and two were in the half-speed group (P22 and P27). Of these five forecasters, tornado warnings verified only for P14, P15, and P22 with associated tornado warning lead times of zero, two, and six min, respectively (Fig. 3.6a).

Sixteen forecasters decided to issue their first (and only) tornado warning reactively, a few minutes after they received the tornado report. Four of the forecasters were in the full-speed group, whereas six were in the half- and quarter-speed groups. The remaining nine forecasters decided not to issue tornado warnings following the report. As most forecasters issued unverified tornado warnings, the median tornado lead time was zero min, and the majority of POD and FAR scores were poor (Table 3.4). In this case, radar update speed had little to no discernable impact on forecasters' performance.

The use of full-speed data was most advantageous during Iota, the case containing a cluster of supercells, one of which produced an EF0-rated tornado (Fig. 3.4b; Table 3.2). In this case, the majority of the full-speed groups' tornado warning lead times were longer than the overall median warning lead time of 0.25 min, which is in stark contrast to the quarter-speed group (Fig. 3.6b). Of the eight in the full-speed group with non-zero tornado warning lead times, half achieved lead times between 25 and 36 min, while the other half achieved lead times under 10 min. Six of 10 participants in the half-speed group achieved non-zero tornado warning lead times; five were five min or less, whereas one was 35 min. The median tornado warning lead times for full-, half-, and quarter-speed groups were 7.5, 3.5, and 0.0 min, respectively. Besides increasing tornado warning lead time, the use of full-speed data in Iota resulted in fewer tornado misses and false alarms (Table 3.4). About 30 min prior to Iota's EF0 tornado, 4.5-in hail and a 61-kt wind were reported (Table 3.2). For these reports, the distributions of severe warning lead times between groups were relatively similar, with a tendency for lower lead times for members of the quarter-speed group (not shown).

Unlike the previous two tornado cases, Zeta presented a classic cyclic supercell that produced several tornadoes, including two rated EF1 and one rated EF2 (Table 3.2). As in the severe case, Delta, of particular interest was whether the use of increasingly rapid updates would enhance the tornado warning lead time for the first tornado occurrence, which in operations tends to be the most difficult to forewarn (e.g., Andra et al. 2002; Brotzge and Erickson 2009). In this case, the full-speed group performed best with about twice as many full-speed participants producing first tornado-warning lead times above the overall median of 12 min (median tornado warning lead time = 14.5 min), compared to the half- and quarter-speed groups (median tornado warning lead times: 9 and 11 min, respectively) (Fig. 3.6c). A few forecasters in the full- and half-speed groups issued tornado warnings with comparatively long lead times ranging from 25–35 min (Fig. 3.6c). These results indicate that the full-speed group and a few forecasters in the half-speed group gained situational awareness unavailable in the 4-min volume updates used by the quarter-speed group. The overall median tornado warning lead times for the second and third tornadoes were similar: 16.5 and 17.5 min, respectively (Fig. 3.6c). Also similar were the lead time distributions associated with these warnings, with a slight tendency for lower lead times for the half-speed group. Regardless of the observed differences in tornado warning lead times between groups, no unverified tornado warnings were issued (Table 3.4).

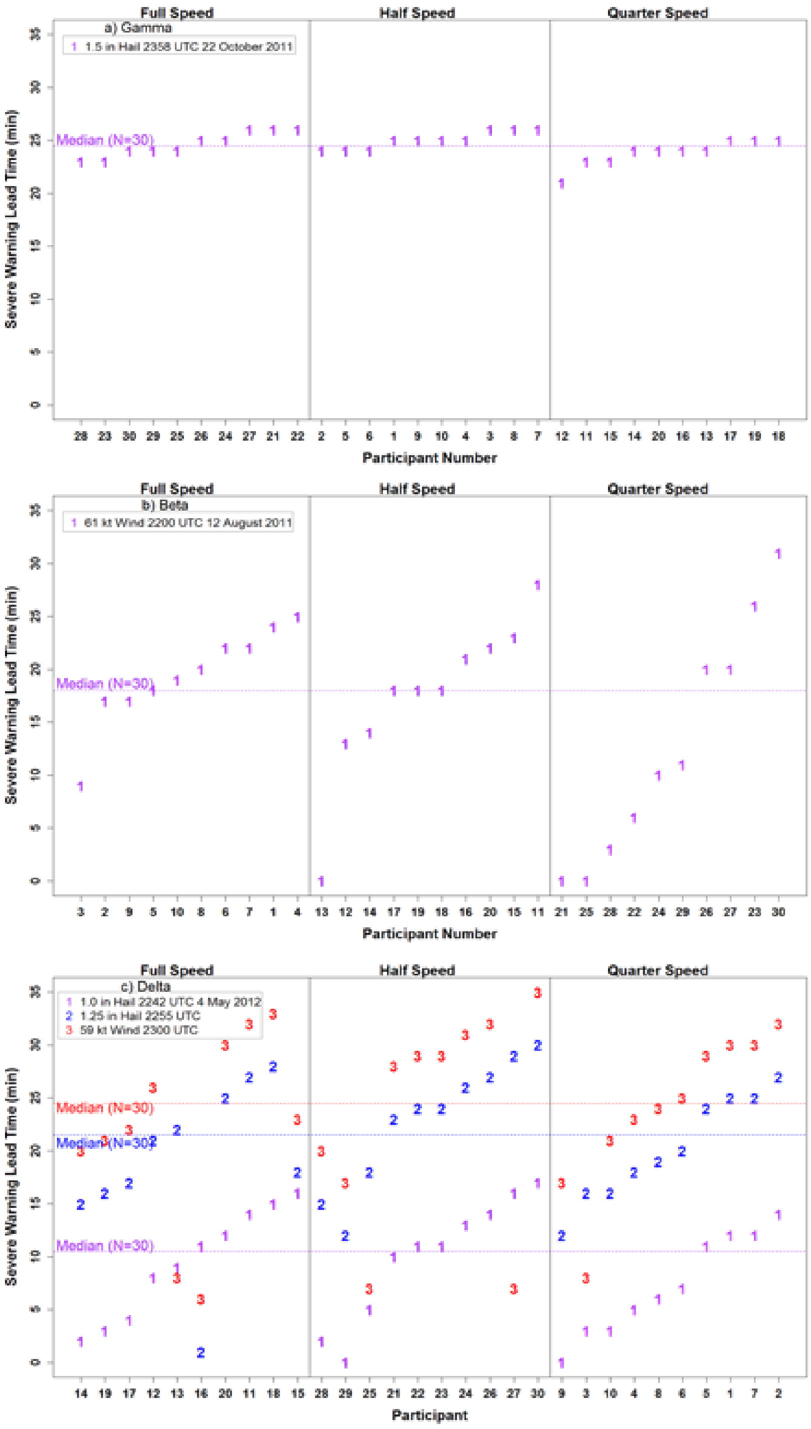


Figure 3.5. Distribution of forecasters' severe warning lead times (min) for each case: a) Gamma, b) Beta, and c) Delta, organized by update speed. First, second, and third severe reports are denoted by numbers 1, 2, and 3 (magenta, blue, and red). For each report, the median severe warning lead time (min) for the full distribution is given by a dotted and annotated line (magenta, red, and blue).

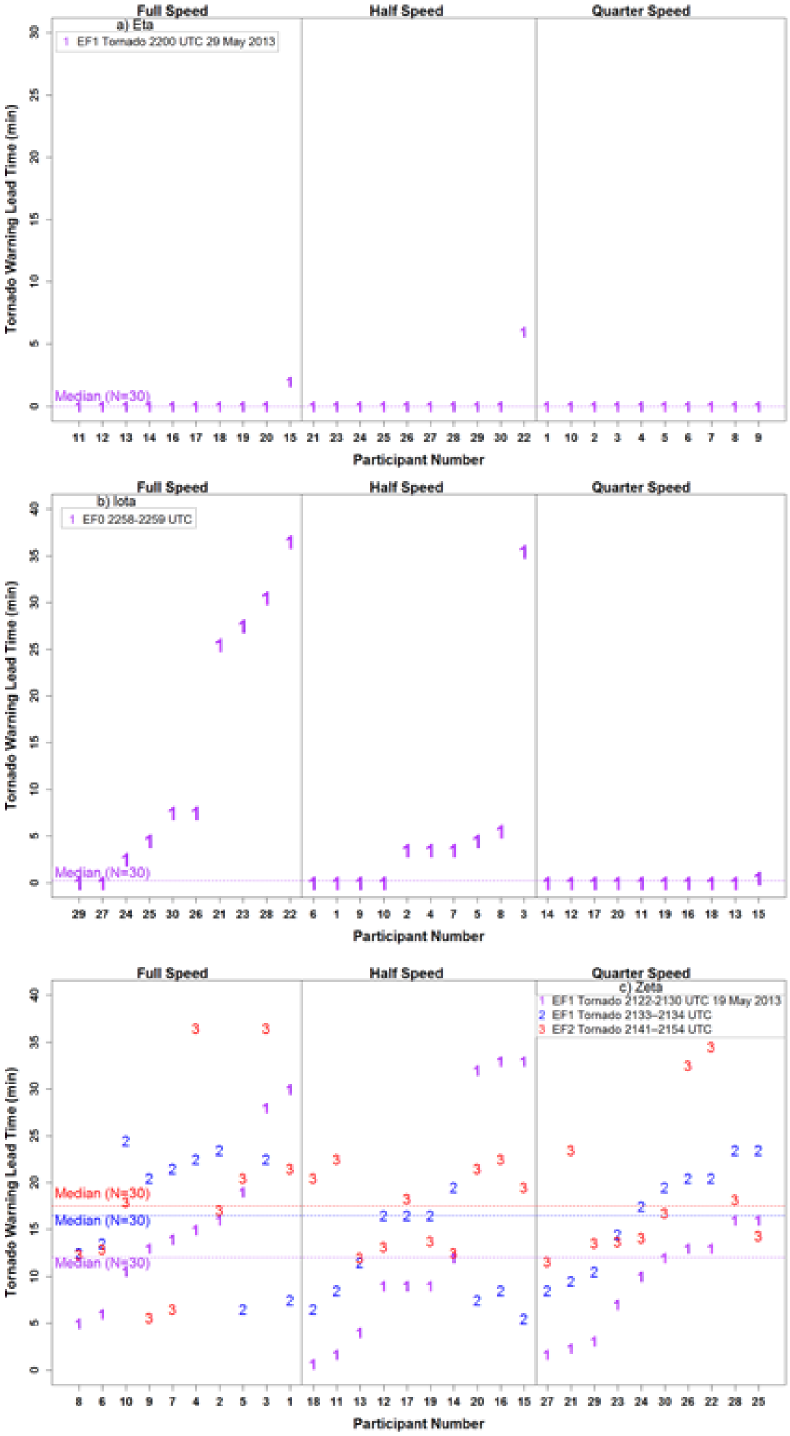


Figure 3.6. Distribution of forecasters' tornado warning lead times (min) for each case: a) Eta, b) Iota, and c) Zeta, organized by update speed. First, second, and third severe reports are denoted by numbers 1, 2, and 3 (magenta, blue, and red). For each report, the median severe warning lead time (min) for the full distribution is given by a dotted and annotated line (magenta, red, and blue).

3.5.3 Performance: Null Cases

Epsilon and Alpha presented forecasters with null multicell events (Fig. 3.2a,b; Table 3.2). Of the two cases, the results indicate that the use of full-speed data was most advantageous during Epsilon, as only 16 of 30 forecasters decided to issue severe thunderstorm warnings. Of the 16 forecasters who issued warnings, three were in the full-speed group, compared to six and seven in the half- and quarter-speed groups. In contrast, while working Alpha (Fig. 3.2b), most forecasters (26 of 30) decided to issue severe thunderstorm warnings. Of the four that did not issue severe thunderstorm warnings, one each used full- and half-speed data, while two used quarter-speed data.

During Theta (Fig. 3.2c; Table 3.2), the nontornadic supercell case, most forecasters (24 of 30) issued severe thunderstorm warnings, and one third issued tornado warnings. To assess severe and tornado warning false alarms separately, the FAR was computed with respect to each warning type (Fig. 3.7). Although the distribution of severe thunderstorm warning FAR scores is fairly similar across update speeds, a few more forecasters achieved severe FAR scores lower than 0.5 using quarter-speed data (N=5) than when using full- or half-speed data (N=3). In contrast, more forecasters using quarter- and half-speed data issued tornado warnings (N=5 and N=4, respectively) than those using full-speed data (N=1). Hence, in this case, the use of full-speed data appeared to be most advantageous in reducing the number of tornado false alarms.

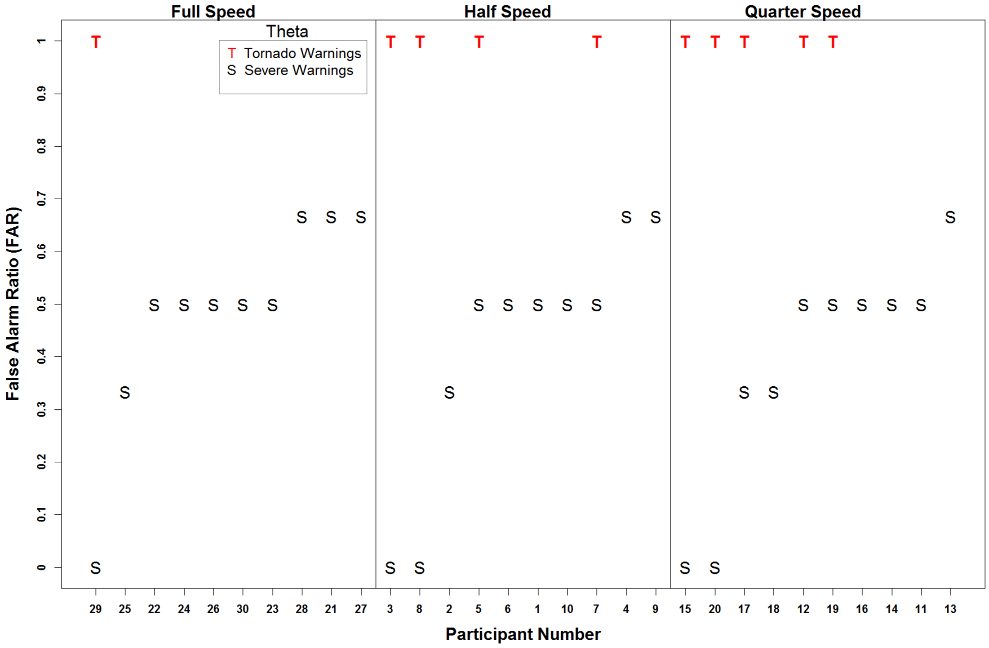
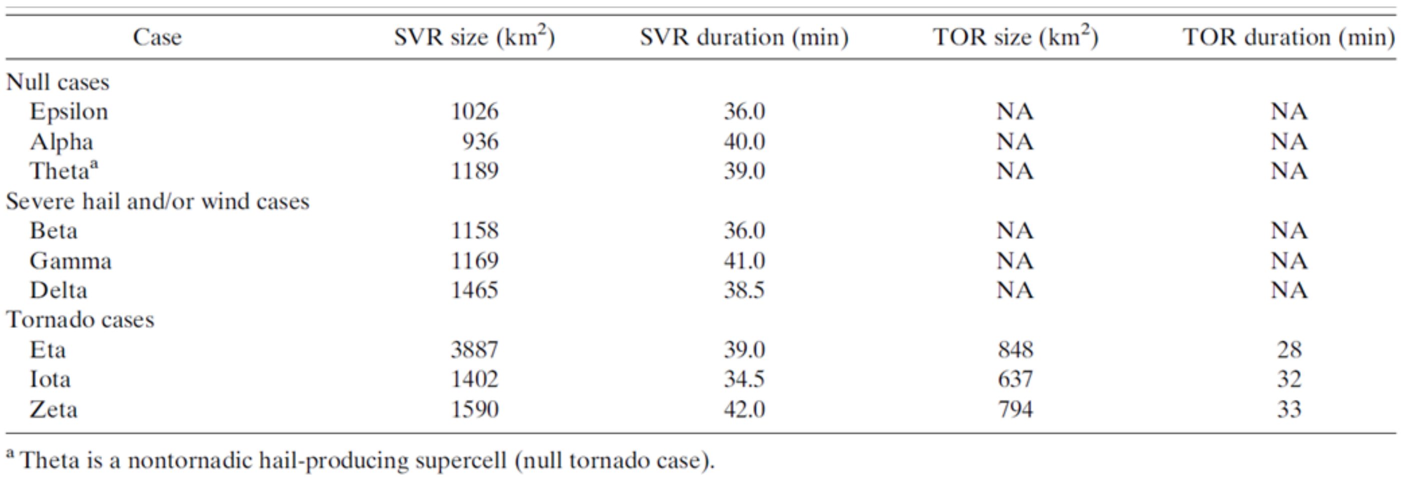


Figure 3.7. The False Alarm Ratio (FAR) for severe (black S) and tornado (red T) warnings by participant during Theta, plotted by update-speed group.

3.6 Warning Polygon Size and Duration

While analyzing forecaster performance, multiple questions arose about whether warning characteristics (i.e., size and duration) depended on storm mode or radar-update speed. We found that the largest differences in warning characteristics were related to each case's storm mode. For example, the largest severe thunderstorm warnings were issued during the squall-line case (Eta; Table 3.5, Fig. 3.8a), which is not surprising given that squall lines can stretch over 100 km in length and can produce widespread severe weather (e.g., Funk et al. 1999; Trapp et al. 2005). Various warning strategies employed by 2015 PARISE participants likely resulted in these very large severe thunderstorm warnings. For example, P15 explained the need for a large warning size during Eta. They stated that their main objective was to warn for the deepest reflectivity core, but that the warning should also capture new deep reflectivity cores and potential severe-weather threats that might develop anywhere along the line.

Table 3.5. Overall median warning size and duration for all warnings issued in each case. Three case types are null, severe hail and/or wind, and tornadic.



Case	SVR size (km^2)	SVR duration (min)	TOR size (km^2)	TOR duration (min)
Null cases				
Epsilon	1026	36.0	NA	NA
Alpha	936	40.0	NA	NA
Theta[a]	1189	39.0	NA	NA
Severe hail and/or wind cases				
Beta	1158	36.0	NA	NA
Gamma	1169	41.0	NA	NA
Delta	1465	38.5	NA	NA
Tornado cases				
Eta	3887	39.0	848	28
Iota	1402	34.5	637	32
Zeta	1590	42.0	794	33

[a] Theta is a nontornadic hail-producing supercell (null tornado case).

Tornado warning size and duration also varied most based on storm mode. Tornado warnings issued during the squall-line case (Eta) were the largest, but the duration of these warnings was the shortest of the three tornado cases (Table 3.5, Fig. 3.9). While working Eta, 12 participants expressed uncertainty in issuing a tornado warning based on radar data alone. In total, 18 participants issued a tornado warning only after receiving a tornado report. Based on performance (section 3.5), Eta was a challenging case and the higher uncertainty expressed by the participants likely influenced the size and duration of their warnings. In addition, 11 participants explicitly stated that squall line tornadoes tend to be short-lived, which likely resulted in shorter-duration tornado warnings. The participants' perception that squall-line tornadoes tend to be short-lived was accurate for this case, as the tornado in Eta lasted one minute (Table 3.2). Studies of tornadoes relative to storm mode also align with the participants' perceptions (e.g., Trapp et al. 2005; Davis and Parker 2014). During the other two tornado cases, environmental conditions alerted participants to a heightened potential for strong supercells that can produce long-lived tornadoes, thereby requiring longer warnings. The classic supercell case (Zeta) had the longest tornado warnings, although these warnings were only one minute longer than those issued during the supercell cluster case (Iota; Table 3.5). During Zeta, participants also received multiple tornado reports throughout the case. Knowledge of a confirmed tornado may explain why 17 of the 30 participants issued a second tornado warning that was longer than the first tornado warning.

While differences in warning size and duration were observed for cases with differing storm modes, it is worth noting that these characteristics did not change substantially when radar update speed changed. In addition, when looking at the cases individually, no clear patterns emerged in terms of warning characteristics and radar update speed (Figs. 3.8, 3.9). Since radar update speed did affect lead time (section 3.5) but not warning characteristics, it is possible that changes in radar update speed affects when, not how, a forecaster designs and issues a warning.

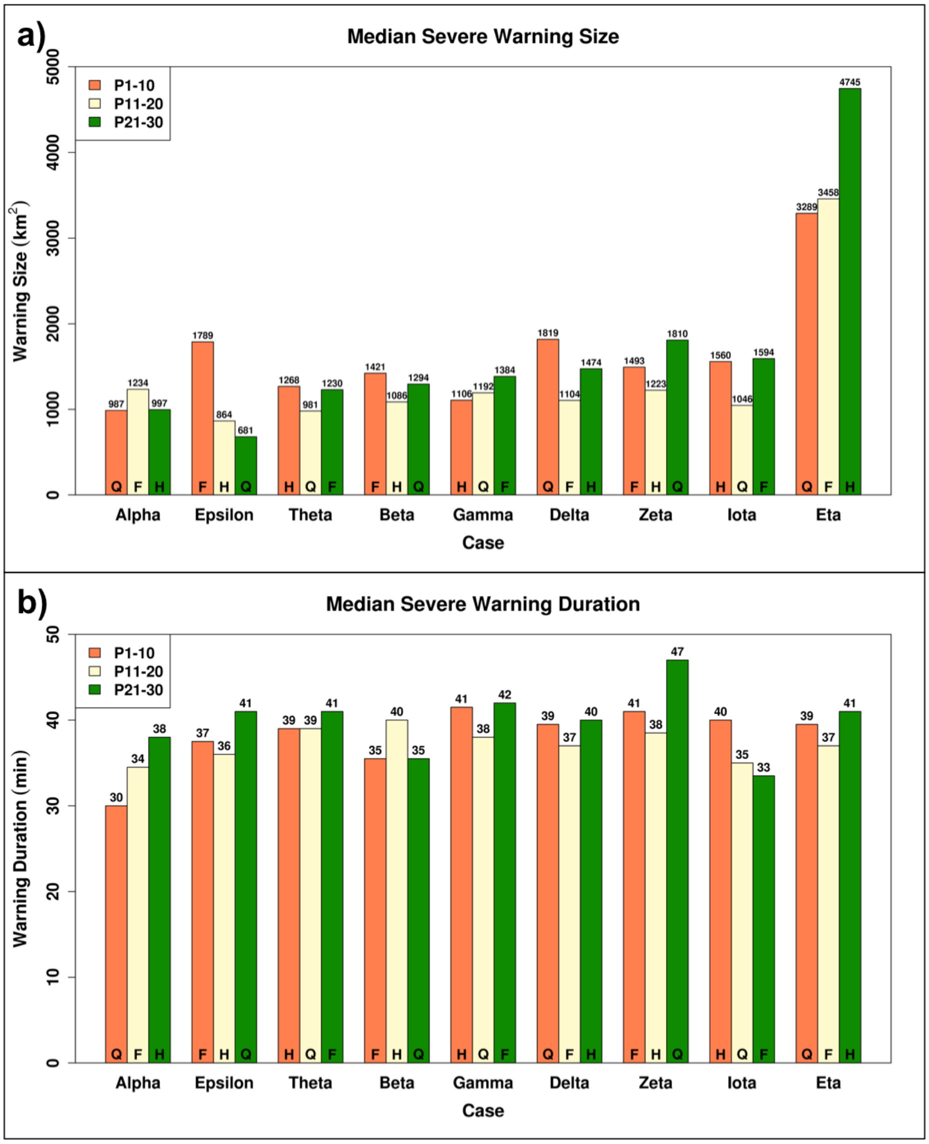


Figure 3.8. Median severe thunderstorm warning a) size and b) duration for each participant group. Median values are included near the top of each bar. Radar-update speed (F=Full, H=Half, and Q=Quarter) worked by each group for each case is included near the bottom of each bar.

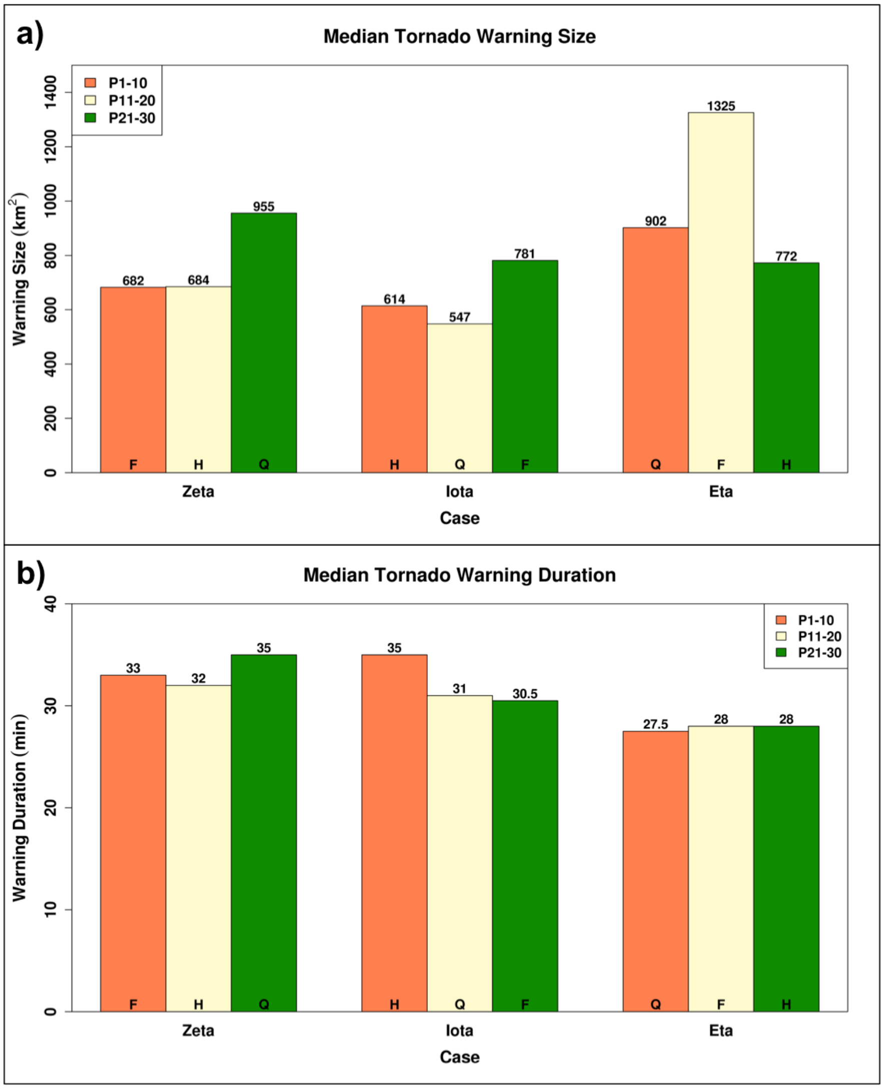


Figure 3.9. Median tornado warning a) size and b) duration for each participant group. Median values are included near the top of each bar. Radar-update speed (F=Full, H=Half, and Q=Quarter) worked by each group for each case is included near the bottom of each bar.

3.7 Cognitive Workload

3.7.1 Workload Distributions and Profiles

The ISA workload analysis is based on forecasters' ratings chosen at 5-min intervals during the video-cued retrospective recall. The number of ratings in each case ranged from 6–13 depending on case duration. In total, 24 ISA ratings were missed, 8 of which each belonged to the quarter-, half-, and full-speed groups. Over half of these missed ratings occurred during the tornado cases, possibly due to the higher demand of this case type. Given that these workload reports were incomplete, they were removed from the analysis.

Each group's median 5-min workload rating for the nine cases was either a level 2 or a level 3 (Fig. 3.10). This result suggests that on average, forecasters were not cognitively overloaded during this experiment. However, a difference in cognitive workload based on temporal resolution is evident. While the quarter-speed group was on average a level 2 (relaxed) for all of the null and severe hail/wind cases, the full-speed group was a level 3 (comfortable) for half of these (Fig. 3.10a-f). The half-speed group was a level 3 for only one of these cases (Theta), which although classified as null, presented a nontornadic supercell that produced severe hail. The median workload rating for the tornado cases was a level 3 for all groups (Fig. 3.10g-i), suggesting that aside from temporal resolution, the increased weather threat contributed to the overall higher levels of workload.

Despite some similarities in the median 5-min workload ratings, a Kruskal-Wallis test (Kruskal and Wallis 1952) showed statistically significant differences in ISA ratings between the three groups in all but two cases (*p*-values<0.05; Table 6). One of these cases, Gamma, was when forecasters' performance was most similar (Fig. 3.5a). A post-hoc Wilcoxon-Mann-Whitney rank-sum test (e.g., Wilks 2006) indicates between which groups these statistically significant differences occurred (*p*-value<0.017; Table 6). Comparing the three groups, the quarter-speed group's ISA rating distribution was most different to that of the full-speed group's, while the half- and full-speed groups' ISA rating distributions were most similar (Table 3.6).

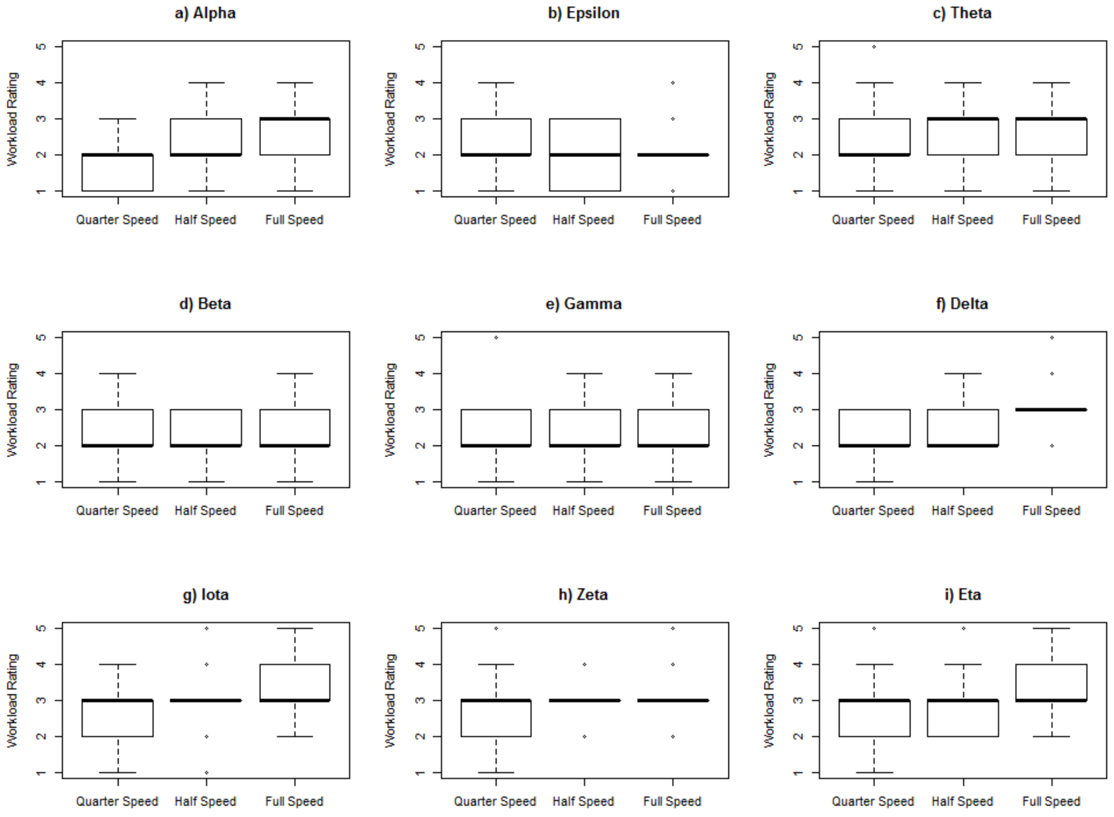
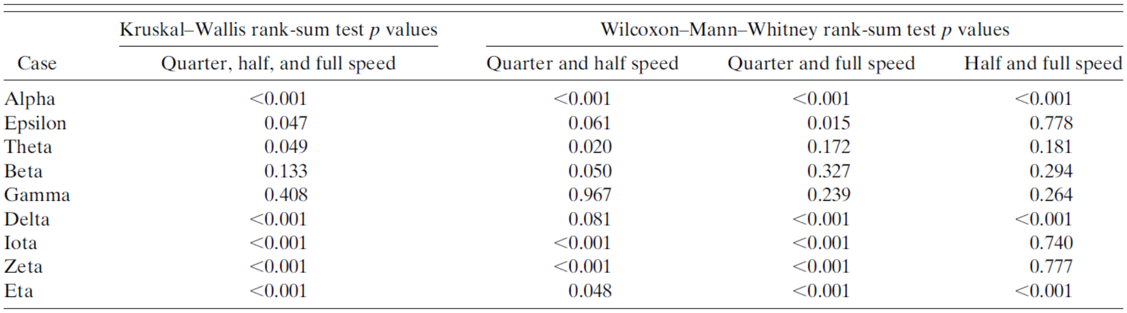


Figure 3.10. Boxplots of 5-min workload ratings for quarter-, half-, and full-speed groups for the null (a, b, c), severe (d, e, f), and tornadic (g, h, i) cases. The solid middle line indicates the median value and the box edges indicate the lower and upper quartiles (i.e., interquartile range). Minimum and maximum values are identified with whiskers and outliers are either less than 1.5 times the lower quartile or greater than 1.5 times the upper quartile.

Comparisons of ISA rating distributions give an overall impression for the level of cognitive workload experienced within a case. However, given the dynamic nature of weather, the change in workload as cases evolved (i.e., workload profile) was also of interest. We observed that regardless of temporal resolution or case type, 21 of the 30 participants' workload rating patterns were either flat (i.e., little or no change in workload) or fluctuating (i.e., multiple increases and decreases in workload) in the majority of cases worked. Although we did not analyze personality traits during PARISE 2015, these workload behavior tendencies suggest that forecaster personality was also likely an important factor in perceived cognitive workload during the simulations. It is possible that personality traits may have influenced forecasters' coping strategies and approaches to the simulations, thus influencing their ISA ratings. Past studies support this suggestion; personality traits and perceived subjective workload have been found to correlate during vigilance tasks (e.g., Rose et al. 2002; Szalama 2002; Guastello et al. 2015). The influence of personality would also explain differences in forecasters' level of boredom versus excitement during cases and why some forecasters were more sensitive to changes in task demand than others.

Table 3.6. Kruskal-Wallis rank-sum test and Wilcoxon-Mann-Whitney rank-sum test p-values for differences in cognitive workload distributions across groups with differing temporal resolution.



	Kruskal–Wallis rank-sum test *p* values	Wilcoxon–Mann–Whitney rank-sum test *p* values		
Case	Quarter, half, and full speed	Quarter and half speed	Quarter and full speed	Half and full speed
Alpha	<0.001	<0.001	<0.001	<0.001
Epsilon	0.047	0.061	0.015	0.778
Theta	0.049	0.020	0.172	0.181
Beta	0.133	0.050	0.327	0.294
Gamma	0.408	0.967	0.239	0.264
Delta	<0.001	0.081	<0.001	<0.001
Iota	<0.001	<0.001	<0.001	0.740
Zeta	<0.001	<0.001	<0.001	0.777
Eta	<0.001	0.048	<0.001	<0.001

3.7.2 Reasoning for Higher Levels of Cognitive Workload

3.7.2.1 Categories

Forecasters' reasoning associated with each ISA rating gives insight into the chosen ratings for perceived cognitive workload. Although the average ISA ratings show that forecasters were generally relaxed and comfortable during the nine cases, many ISA ratings extended to a level 4 (high workload), and there are numerous outliers rated at a level 5 (excessive workload) (Fig. 3.10). The reasoning provided for all level 4 and level 5 ISA ratings were analyzed (N=183), and six categories were identified. In order of prevalence, these categories are 1) storm characteristics, 2) warnings, 3) case startup, 4) temporal resolution, 5) technical frustrations, and 6) personal (Fig. 3.11a). Storm characteristics causing higher cognitive workload included the number of storms in the sector, expected threat, and evidence of intensification. The warning category is associated with higher cognitive workload due to the extra task of issuing products, sacrificing interrogation time, having concern about polygon placement relative to storms, and the unfortunate realization that warnings were not panning out as expected. Case startup describes increased workload that was experienced within the first 5–10 min of a case. During this time, higher cognitive workload was experienced because forecasters felt an urgency to load their data, assess the situation, and possibly make warning decisions. The temporal resolution of radar data was associated with higher workload, such that forecasters felt the need to monitor the data quickly so that they could keep up with trends. Oftentimes forecasters reported higher levels of workload because they did not have enough time to look at all the data and were not able to pinpoint the important signals. Technical frustrations caused increases in workload typically because WarnGen/AWIPS-2 did not function as it should, which sometimes caused delays in product issuance. Finally, one forecaster reported three ISA ratings of level 5 due to requiring a bathroom break while monitoring the weather.

3.7.2.2 Temporal Resolution

Forecasters using full-speed PAR data reported approximately twice as many level 4 and 5 ISA ratings than those using quarter- and half-speed PAR data (Fig. 3.11b). The largest reasoning category for the full-speed group's higher ISA ratings was storm characteristics, followed by temporal resolution (Fig. 3.11b). In comparison, only a small portion of the half-speed participants reported higher ISA ratings due to temporal resolution, and no quarter-speed participants' reasoning related to temporal resolution (Fig. 3.11b). Storm characteristics and warning categories accounted for more than half of the reasoning for the quarter- and half-speed groups (Fig. 3.11b). Technical frustrations also accounted for a large portion of the quarter-speed group's higher ISA ratings, while case startup accounted for a quarter of the half-speed group's (Fig. 3.11b).

Only a small fraction of the higher cognitive workload ratings were a level 5 (N=26). However, these ratings cause most concern because they describe a mental state that is cognitively overloaded. Forecasters using full-speed data gave over half of these ratings (N=16), and related these ratings to every category except for technical frustrations. In comparison, almost all of the level 5 ratings given by quarter-speed participants were due to technical frustrations (N=5 of 7). The remaining level 5 ratings given by quarter- and half-speed participants were associated with case startup and warning reasoning. Excessive workload due to temporal resolution, storm characteristics, and personal matters only occurred with full-speed participants.

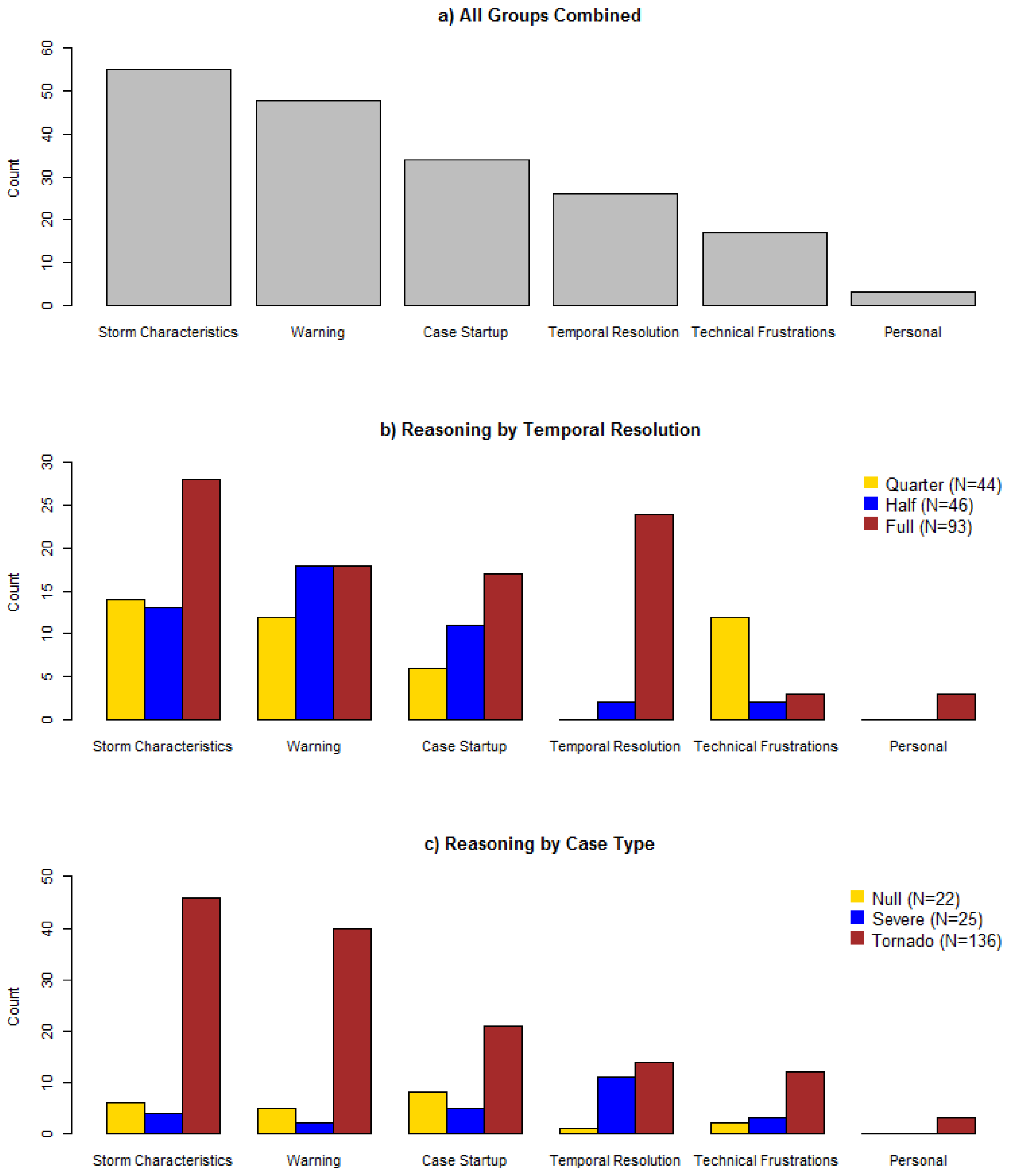


Figure 3.11. Reasoning categories for ISA ratings given at levels 4 and 5 for a) all groups combined, b) each temporal resolution, and c) each case type.

3.7.2.3 Storm Type

Of all the case types, forecasters reporting level 4 and level 5 ISA ratings did so most during the tornado cases (Fig. 3.11c). Reasoning for this increase in cognitive workload was mostly associated with the storm characteristics and warning categories. Monitoring multiple threats for one supercell, dealing with uncertainty in storm evolution, and feeling overwhelmed with the number of warning products needing to be issued were all factors leading to these higher levels of experienced cognitive workload. Although temporal resolution was not a large contributor to the higher cognitive workload reported during the tornado cases, it was the largest category for why forecasters reported these higher ISA ratings during the severe hail/wind cases (Fig. 3.11c). The temporal resolution reasoning was mostly associated with Delta, and occurred due to forecasters not being able to examine the data closely as updates were coming in, having difficulty comprehending the structure and evolution of the storm due to the fast updates, and needing to adapt to a different type of interrogation strategy. It is worth noting that update speeds were quickest in Delta compared to the other cases (Table 3.2). The different reasoning driving level 4 and 5 ISA ratings for tornado and severe hail/wind cases supports that higher cognitive workload is not only a function of temporal resolution, but also of storm type, as suggested earlier.

3.8 Discussion

Based on the performance analysis, we found that forecasters' ability to increase severe and tornado warning lead times when using increasingly higher-temporal resolution data depended on the weather situation presented. Distributions of positive warning lead times were most comparable during Gamma (Fig. 3.5a); this result suggests that similar situational awareness was gained by forecasters in all three groups. While working the two other severe cases, the use of increasingly higher-temporal resolution data most aided forecasters' ability to issue verified warnings earlier during Beta, the severe wind event (Fig. 3.5b). A tendency for longer initial warning lead times when using increasingly higher-temporal resolution data was also found during Delta, the hail and wind event (Fig. 3.5c). These findings are consistent with Bowden et al. (2015), who in PARISE 2013 found the use of full-speed PAR data, compared to quarter-speed PAR data, increased median severe thunderstorm warning lead times by 5 min in two severe (large hail and/or damaging wind) cases. In a follow-on study by Bowden and Heinselman (2016), their analyses of forecasters' situational awareness determined that longer severe thunderstorm warning lead times were driven by forecasters' ability to observe rapid changes in radar-based hail and wind precursors earlier when using 1-min vs 5-min radar volume scans. More frequent sampling of specific hail and wind events by PAR was also found to improve scientific understanding of radar-based severe storm precursors in several case studies, including Heinselman et al. (2008), Emersic et al. (2011), Newman and Heinselman (2012), and Kuster et al. (2016). The advantage of frequent updates in the analysis of severe storms, and in particular downbursts, has been demonstrated in prior studies using rapid-scan data from other radar platforms (e.g., Roberts and Wilson 1989).

This PARISE was the first in the series of former experiments to explore the ability of forecasters to issue verified tornado warnings with lead time in advance of a short-lived tornado within a bowing line segment. During this event (Eta), the overall lack of verified tornado warnings with positive lead time, especially when using full-speed data, is somewhat discouraging (Fig. 3.6a). Our expectation for a more positive result was supported by the regional radar climatology of tornadic and nontornadic vortices within nonsupercell storms by Davis and Parker (2014), who found statistically significant differences in their azimuthal shear magnitudes ($0.006s^{-1}$ or higher) when located within 60 km of a WSR-88D. The velocity couplet associated with the Eta tornado was located 15 km outside of this ideal radar range. Davis and Parker (2014) also found the median detection lead time for these nonsupercell tornadic vortices was 10 min, which suggests that the use of 1- or 2-min volume updates has the potential to improve forecasters' detection lead time for such events. While future analyses of participants' retrospective data will provide insight into this finding, anecdotal conversations with NWS forecasters reveal that some forecasters either do not issue tornado warnings during these types of events or wait for confirmation of a first event, owing to the potential for high false alarm rates. Additionally, when bowing lines (like this one) are fast moving, some forecasters discern the impact of the storm's translational motion as a more significant threat than the embedded circulation, and therefore issue severe thunderstorm warnings instead.

In contrast, for the two tornadic supercell cases (Zeta and Iota), forecasters' ability to issue verified and timely tornado warnings on the first tornado event improved when using full- and half-speed PAR data (Fig. 3.6). Zeta, a "classic" tornadic supercell event, appeared to be the more straight-forward event since all issued tornado warnings verified. Iota, a tornadic supercell cluster, appeared more challenging, as full- or half-speed data were needed to achieve verified tornado warnings with lead time. Additionally, during the nontornadic supercell case (Theta), the use of full-speed data aided forecasters' ability to discriminate correctly the severe weather threat, resulting in fewer false alarms (Fig. 3.7). Together these results are consistent with the 2010 and 2012 PARISE findings of Heinselman et al. (2012, 2015), where the use of higher-temporal resolution also resulted in longer tornado warning lead times. However, FAR results were mixed, as FAR was impacted negatively in PARISE 2010 and positively in PARISE 2012 when using faster radar updates (Heinselman et al. 2012, 2015, respectively). The PARISE 2015 FAR results are most consistent with the PARISE 2012 FAR findings. The advantage of frequent updates in the analysis of a potentially tornadic supercell's storm evolution, including specificity of tornado movement, has been demonstrated in prior studies using PAR data (e.g., Kuster et al. 2015) as well as data from other weather radars (e.g., Vasiloff 2001; Wurman et al. 2012; e.g., Isom et al. 2013; Pazmany et al. 2013; Kurdzo et al. 2015).

3.9 Conclusions and Future Work

The purpose of this paper was to focus on the traditional experiment component of the 2015 PARISE and share performance, warning characteristics, and cognitive workload results. The increased number of participants and cases worked compared to earlier experiments improves the generalizability of our work. The overall finding that median warning lead time increased with increasing update speed is in line with our findings from previous studies. Earlier warnings were provided in two severe hail/wind and two tornado cases, and the use of full-speed data for discriminating the weather threat was particularly useful to forecasters during Theta. However, longer warning lead time with faster update speeds was not observed in all cases, most notably during Eta. This finding suggests that specific training and guidance may be required to fully realize the benefits of full-speed PAR data to forecasters' warning decision processes during more challenging events. Making use of dynamic scanning methods that are already available (e.g., Chrisman et al. 2009, 2014) will be a helpful first step to developing the skills necessary for processing rapidly-updating radar data during warning operations.

While the update speed impacted when warnings were issued, it did not influence the size or duration of warning polygons (Figs. 3.8, 3.9). Therefore, further improvements to warning metrics (such as the false alarm area) may require a change in the warning paradigm. This change may be possible through modernization of the current NWS warning system. A move towards probabilistic hazard information via the FACETs (Forecasting a Continuum of Environmental Threats) framework is expected to address multiple aspects of warning characteristics (e.g., Stumpf et al. 2008 and Karstens et al. 2015).

Forecasters' subjective assessments of cognitive workload within the PARISE setting suggest that cognitive workload will rarely reach excessive, and when it does, it could be due to a variety of reasons that are not necessarily tied to the temporal resolution of radar data. Our data also suggests that perceived cognitive workload may relate to forecasters' personality. Although we have not yet explored this relationship scientifically, investigating this hypothesis would be beneficial to a number of testbed experiments that may also observe effects of individual differences on forecasters' approaches, performance, and perceived workload.

Despite increasing our sample size and the variety of cases worked, we must be mindful of the limitations that still remain in this experiment. In these simulations, forecasters' warning decision processes were isolated to their independent thought; unlike in the forecast office, forecasters did not work in teams and therefore the data collected is not an accurate reflection of what could be expected in real warning operations. Additionally, forecasters' limited access to radar products and the absence of dual-polarization radar data simplified their warning decision processes even further. Considerations of these missing elements and how a future operational PAR system might impact convective warning operations will be addressed in the PARISE 2015 focus group analysis.

Chapter 4

Exploring Applications of Eye Tracking in Operational Meteorology

Taken in full from: Wilson, K. A., P. L. Heinselman, and Z. Kang, 2016: Exploring applications of eye tracking in operational meteorology research. *Bull. Amer. Meteor. Soc.*, **97**, 2019–2025.

Abstract

Eye-tracking technology can observe where and how someone's eye gaze is directed, and therefore provides information about one's attention and related cognitive processes in real time. The use of eye-tracking methods is evident in a variety of research domains, and has been used on few occasions within the meteorology community. With the goals of Weather Ready Nation in mind, eye-tracking applications in meteorology have so far supported the need to address how people interpret meteorological information through televised forecasts and graphics. However, eye-tracking has not yet been applied to learning about forecaster behavior and decision processes. In this article, we consider what current methods are being used to study forecasters and why we believe eye-tracking is a method that should be incorporated into our efforts. We share our first data collection of an NWS forecaster's eye gaze data, and explore the types of information that this data provides about the forecaster's cognitive processes. We also discuss how eye-tracking methods could be applied to other aspects of operational meteorology research in the future and provide motivation for further exploration on this topic.

4.1 Background

Recently, eye-tracking has been used within the meteorology community to assess communications of weather information to the public. Drost et al. (2013) used eye-tracking to study the impact of a weathercaster's gesturing during a televised weather forecast on viewers' attention. Their analysis revealed that while gesturing did not impact viewers' retention of information, it did redirect viewers' attention to different elements on the screen. Eye-tracking was also used by Sherman-Morris et al. (2015) to investigate the effectiveness of different legend colors and content in hurricane storm surge graphics on participants' ability to accurately interpret threat levels. Although significant differences in accuracy were not found across legends of different color and content, participants' eye-tracking data indicated they struggled most when the legend color was shades of blue and the values were in feet. Studies such as these are helping the United States work towards becoming a Weather Ready Nation. A Weather Ready Nation is one that builds community resilience to increasing vulnerability of extreme weather and water events (NOAA 2015). Lindell and Brooks (2013) summarized a number of major issues that a Weather Ready Nation workshop in 2012 identified as requiring attention. Conducting cognitive research in laboratory experiments to understand users' interpretation of forecasts and warnings was one identified issue (Lindell and Brooks 2013). The studies described by Drost et al. (2013) and Sherman-Morris et al. (2015) demonstrate ways in which eye-tracking is being used to help address this issue.

Another issue identified in the 2012 Weather Ready Nation workshop was the need to study forecasters through behavioral research (Lindell and Brooks 2013). Highlighted was the need for research to develop an understanding of forecasters' decision making processes and how they differ between individuals and the NWS regions. To date, forecaster decision making processes have been examined using a variety of qualitative methods. For example, an ethnographic approach was used by Daipha (2015) to observe and study how forecasters collect and use information in the forecast office. Root Cause Analysis is also performed by forecasters after an event has occurred so that they can evaluate their own warning decisions (Quoetone 2009). Root Cause Analysis encourages forecasters to reflect on their decision making processes and helps uncover reasons for why problems occur. The Critical Incident Technique has also been used in research to gather stories of forecasters' descriptions of past events and what their associated behaviors were (LaDue et al. 2010). Furthermore, research in the NOAA Hazardous Weather Testbed has used surveys and blogs to collect forecasters' feedback of new products tested during warning operations (Calhoun et al. 2014). A retrospective recall method has also been used in the Hazardous Weather Testbed to study individual forecaster's cognition associated with radar data interrogation (Heinselman et al. 2015 and Bowden et al. 2015). This method collects video-cued recall information while forecasters watch a playback video of their onscreen activity and verbalize their past thought processes. Specifically, this method yields detailed information about what forecasters see, think, and do while interrogating radar data. Although retrospective recall data have been incredibly insightful, the complexity of forecasters' decision processes means that the use of qualitative methods alone do not fully capture the intricate cognitive processes of forecasters.

To our knowledge, eye-tracking has not been applied to study NWS forecasters' decision making and related cognitive processes. However, applications of eye-tracking in a variety of research domains, including the studies carried out by Drost et al. (2013) and Sherman-Morris et al. (2015), suggest that this tool could enrich our understanding of how forecasters use information to make decisions. Studies in research domains such as air traffic control and medicine demonstrate how eye-tracking can be used to ask questions that—in an analogous sense—we may wish to answer in operational meteorology. For example, Kang and Landry (2014) used eye-tracking to analyze how novice and expert air traffic controllers' eyes scanned a radar display during aircraft conflict detection tasks. Kang and Landry (2014) found that training novices with experts' scanpaths reduced novices' number of false alarms. We may wonder in operational meteorology how low- and high-performing forecasters' scanpaths of weather radar data differs, and whether such information may be helpful during training. Wood et al.'s (2013) study on visual expertise of radiologists during detection and diagnosis of skeletal fractures is also relatable to operational meteorology. After all, forecasters use radar data to detect the potential for severe weather and then correctly diagnose what type of threat they expect. In Wood et al.'s (2013) study, radiologists' eye gaze data were used to measure their accuracy and speed, which are also measures used to analyze forecaster performance.

4.2 Example: Understanding a Forecaster's Decision Process

To explore how forecasters' eye gaze data may enrich our current understanding of their decision processes, we collected an NWS forecaster's eye gaze data as he interrogated radar data from one weather event, and subsequently obtained his retrospective recall. Eye-tracking research is built on the foundation of the eye-mind hypothesis, such that we assume a person's eye gaze indicates where their attention is and what is at the “top of the stack” of their cognitive processes (Just and Carpenter 1976b). Therefore, measuring forecasters’ eye gaze behavior may provide a way for us to learn about their cognition at a deeper level. The goal of this short study was not to draw conclusions about forecaster cognitive processes, but to think about what type of information eye-tracking methods can provide for learning about cognitive processes that our current qualitative methods do not.

During this short study, the forecaster viewed a 39-min long severe hail and wind event from 16 July 2009 in displaced real time and was asked to make warning decisions as he saw necessary. During this event, a nonsevere northern storm and severe southern storm moved south towards Oklahoma City, Oklahoma. The nonsevere northern storm was well developed at the beginning of the case, while the southern storm was captured from early in its initiation. The forecaster viewed 1-min base velocity and reflectivity PAR updates (Zrnić et al. 2007 and Heinselman and Torres 2011) using the Warning Decision Support System-Integrated Information (WDSS-II; Fig. 4.1). The forecaster was able to loop through radar data, navigate in time and by elevation using function keys, and zoom in and out. Warnings were issued using a polygon tool located in the control panel.

Throughout the simulation, the forecaster’s eye gaze data were collected using the Tobii TX300 eye-tracking system (Fig. 4.1). This system sat below the forecaster’s computer monitor from which an infrared camera detected the location of his pupils and corresponding eye movement on the screen. We viewed the forecaster’s eye gaze data using the Tobii Studio 3.3.0 software, and used a velocity-threshold filter algorithm to identify when and where the forecaster’s eye fixations occurred (Olsen 2002). The forecaster's fixations describe times when his eye gaze momentarily focused on a specific location. The focus is long enough such that he was able to encode and process information (Poole and Ball 2006). The fixation algorithm provided timestamp, duration, and x and y position information for each fixation that the forecaster made. Additionally, we were able to see whether his fixations were made within the reflectivity, velocity, or control panels by creating three separate areas of interest (AOIs; Fig. 4.2a). Defining AOIs in eye-tracking analysis is common practice as this method allows for different types of information presented on the same screen to be distinguished from one another. While the reflectivity and velocity panels presented information about the storms, the control panel provided a polygon tool for issuing warnings.

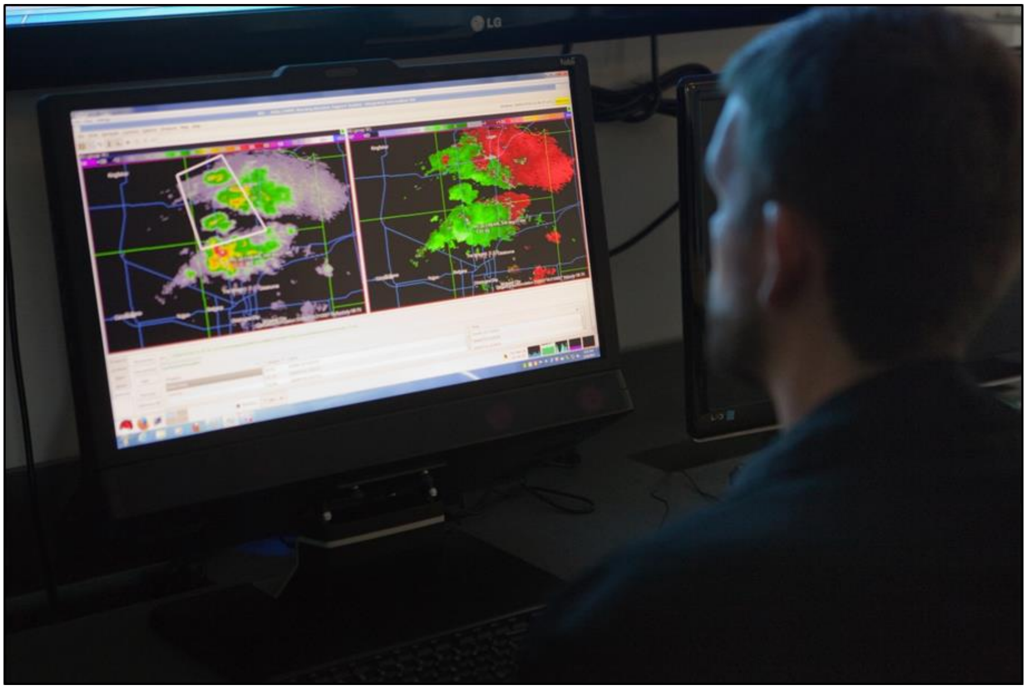


Figure 4.1. Forecaster interrogating 1-min base velocity and reflectivity PAR data using the Warning Decision Support System-Integrated Information. The Tobii TX300 eye-tracker is positioned below the monitor.

We looked at two measures of fixation during this study: fixation count and fixation duration. Higher numbers of fixation count on a particular AOI indicates that the information was either more noticeable or important, whereas longer durations of fixations on a particular AOI indicate that the information was either more engaging or that a greater mental effort was required to extract the information (Poole and Ball 2006). Unlike retrospective recall information, the forecaster's eye gaze data can be used to obtain detailed information about the spatial distribution and temporal trends of these fixation measures in each of the three AOIs. We were interested to see how these fixation measures compared across the three AOIs for the full simulation and how their values changed as the weather scenario evolved. Additionally, we looked at how the forecaster's fixation measures corresponded to the information provided in his retrospective recall, and whether together these two data sets offer a more holistic and accurate understanding of his decision process.

4.3 Counts and Durations of Eye Fixations

Heatmaps are visualizations of the overall spatial distribution of eye fixations within specified AOIs (Fig. 4.2). In Fig. 4.2b, we see that the forecaster fixated most often on the Reflectivity and Velocity AOIs, and least often within the Controls AOI, indicating focus on data interrogation and limited use of the control panel to issue warning polygons (Figs. 4.2b and 4.2c). The distributions of 1-min fixation count and mean duration support this interpretation (Fig. 4.3). Applying the Wilcoxon rank sum test, statistical significance (p<0.05) was established for the difference in median values of 1-min fixation counts and 1-min mean fixation durations across all three AOIs (Figs. 4.3a and 4.3b). Variations in the spatial patterns of total fixation count seen in the Reflectivity and Velocity AOIs suggest the forecaster interrogated these fields differently. A comparison of these heatmaps to the most typical positioning of radar data on the WDSS-II display during the simulation (Fig. 4.2a) indicates that the forecaster fixated nearly equally on the northern and southern storms in the Reflectivity AOI, whereas he fixated more on the southern storm in the Velocity AOI (Figs. 4.2a and 4.2b). In Fig. 4.2c, we see small pockets of longer absolute fixation duration focused on the two storms of interest, however these pockets are more evident in the Velocity AOI. These pockets of longer absolute fixation duration indicate periods of data interrogation focused on specific radar signatures and that the longest fixation duration was on signatures within the Velocity AOI. Differences in fixation measures between the Reflectivity and Velocity AOIs suggest that the forecaster used reflectivity data to interrogate both storms and maintain situational awareness of weather within the entire sector, whereas his interrogation of the velocity data was more directed and focused on regions of storms that were of greatest interest.

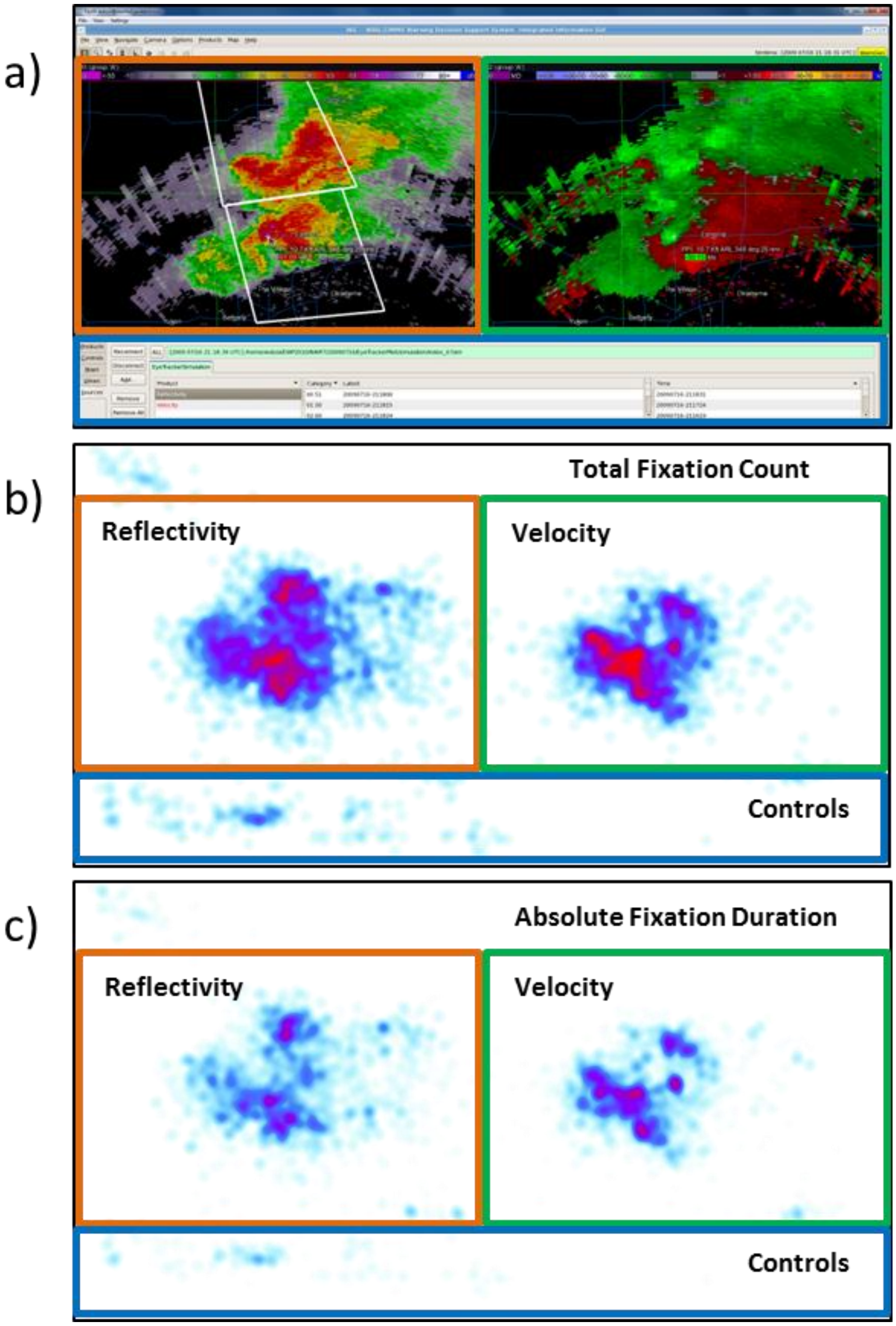


Figure 4.2 a) The Warning Decision Support System-Integrated Information display divided into three areas of interest: reflectivity (left panel, orange box), velocity (right panel, green box), and controls (bottom panel, blue box). Heatmaps were created for the b) total fixation count and c) absolute fixation duration for the entire case. Within the heatmaps, red values indicate a higher fixation count and absolute fixation duration, and blue colors indicate a lower fixation count and shorter absolute fixation duration.

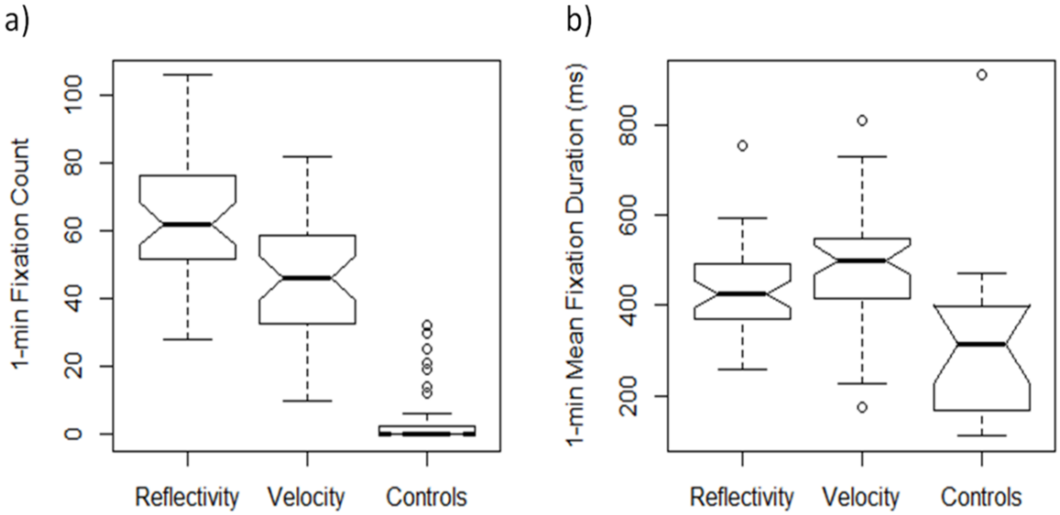


Figure 4.3. Boxplots showing the distribution of a) the 1-min fixation count and b) the 1-min mean fixation duration for the reflectivity, velocity, and controls AOIs. Boxplot whiskers indicate minimum and maximum values, the solid middle line indicates the median value, and lower and upper box edges indicate the interquartile range. Outliers are either less than 3/2 times the lower quartile or greater than 3/2 times the upper quartile. Strong evidence of differing medians is indicated by non-overlapping notches.

4.4 Fixation Trends

Trends in the forecaster's fixation counts and mean fixation durations were seen in the 39-min simulation as the weather scenario unfolded (Fig. 4.4). The interpretation of these trends is aided by computing fixation counts at five-min intervals, resulting in eight periods (with the final period being four min). These trends are of interest because they indicate variations in the forecaster's cognitive activity. While the forecaster fixated most frequently within the Reflectivity AOI, the peak fixation count occurred during the fourth period (Fig. 4.4a). In contrast, the peak fixation count in the Velocity AOI occurred in the seventh period and exceeded the corresponding Reflectivity AOI fixation count.

While in most periods the durations of five-min Velocity AOI fixations were longest, a minimum in Velocity AOI fixation duration occurred in period four when fixation duration and fixation counts in the Reflectivity AOI were longer and higher (Fig. 4.4). Like fixation counts in the Control AOI, the associated fixation durations were intermittent and tended to be shorter than those in the other two AOIs (Fig. 4.4).

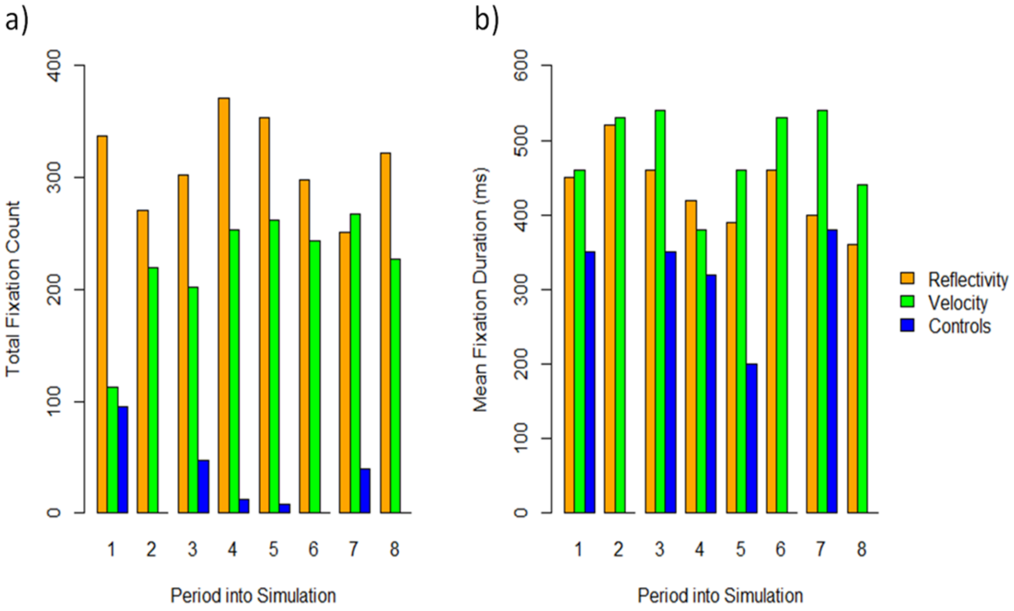


Figure 4.4. a) Total fixation counts and b) mean fixation durations within the reflectivity (orange), velocity (green), and controls (blue) AOIs per period.

To provide context on how these trends in cognitive activity related to different stages of the forecaster's warning decision process, we created a timeline that summarizes the forecaster's retrospective recall during each period (Fig. 4.5). The initial high number of Reflectivity AOI fixation counts in period one resulted from using these data to assess storm intensity. After monitoring trends in the height and intensity of the northern and southern storms' reflectivity cores, the forecaster's decision to issue a severe warning on the northern storm coincided with the highest peak in the Controls AOI fixation count and relatively long fixation durations (Fig. 4.4a). Similarly, the two other peaks in the Controls AOI fixation counts and durations (Fig. 4.4b) coincided with the issuance of severe weather warnings (Fig. 4.5). The increasing trend in Reflectivity AOI fixation count from a relative minimum in period two to its highest peak in period four corresponded with the forecaster's observations of the intensifying southern storm, which he warned on by period three, and by period four he interpreted as being "pretty impressive" with reflectivity values of 70 dBZ up to 25 kft. His focus on reflectivity data also increased because the intensity of the northern storm was diminishing rather than increasing as he had anticipated.

As the southern storm evolved, the downburst potential became apparent to the forecaster and a change in his cognitive process was noticeable in both his fixation trends and retrospective recall. During periods five through seven, the forecaster's fixation count in the Reflectivity AOI decreased and mean fixation duration in the Velocity AOI increased (Figs. 4.4a and 4.4b). Concurrently, he began to observe more interesting signatures in the velocity data (Fig. 4.5). In period five he saw a spatial increase in "downdraft air" in the southern storm as well as the presence of "strong cloud-top divergence" (Fig. 4.5). Although low-level radial winds in the southern storm were only 30–40 kts, the forecaster thought it was "only a matter of time before it really [got] going." His expectation was confirmed in period seven when he saw "intense winds becoming concentrated along the highway." It was also this period that marked the only time that the forecaster's fixation count in the Velocity AOI was higher than in the Reflectivity AOI, and the mean fixation duration in the Velocity AOI was at a maximum. Following from his observation in the velocity data, he decided to issue a second warning on the southern storm, which corresponds with the third peak in fixation count for the Controls AOI (Fig. 4.4a).

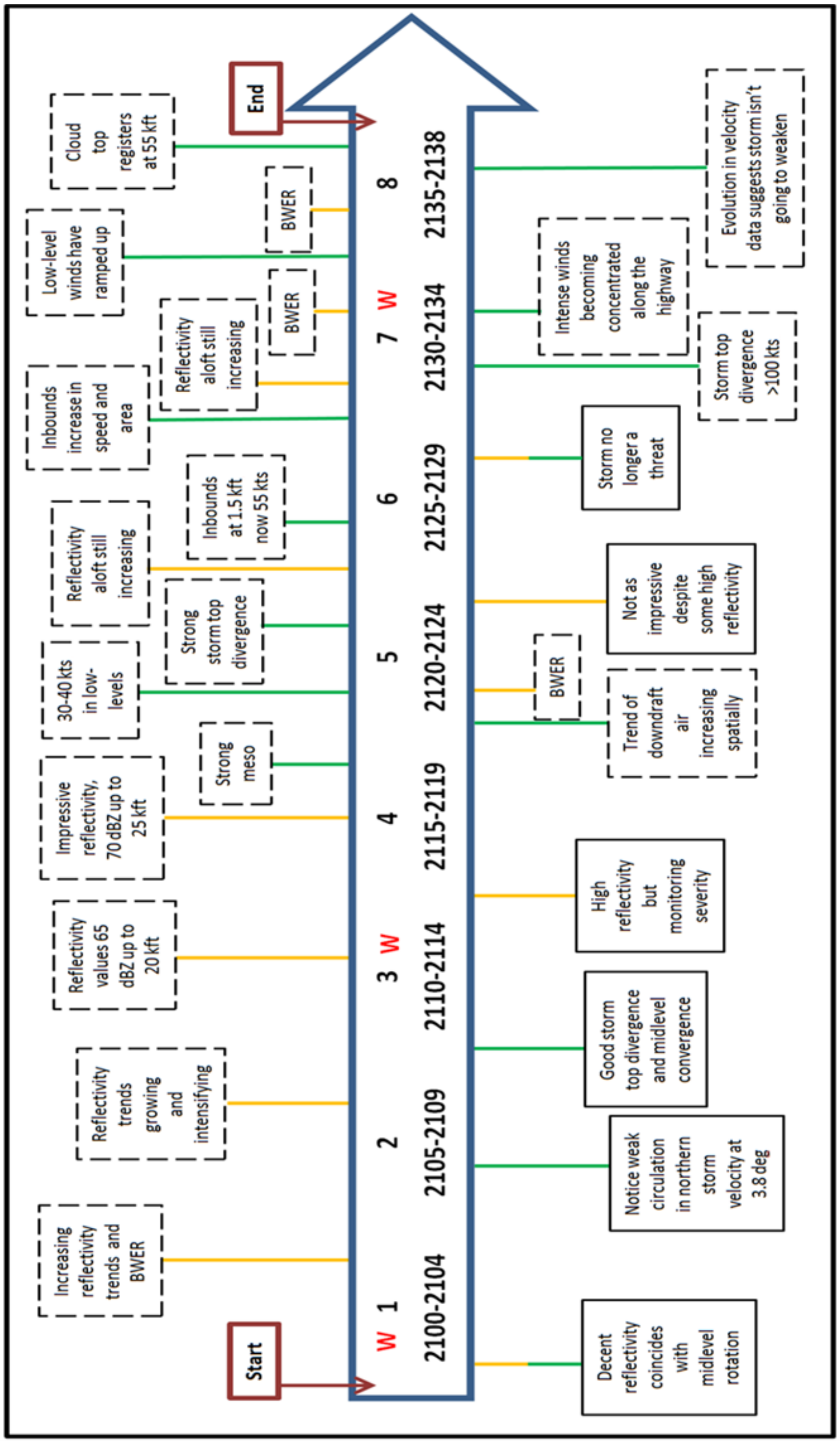

Start
W 1
2100-2104
Increasing reflectivity trends and BWER
Decent reflectivity coincides with midlevel rotation
2
2105-2109
Reflectivity trends growing and intensifying
Notice weak circulation in northern storm velocity at 3.8 deg
W 3
2110-2114
Reflectivity values 65 dBZ up to 20 kft
Good storm top divergence and midlevel convergence
High reflectivity but monitoring severity
4
2115-2119
Impressive reflectivity, 70 dBZ up to 25 kft
Strong meso
Trend of downdraft air increasing spatially
5
2120-2124
30-40 kts in low-levels
Strong storm top divergence
BWER
Not as impressive despite some high reflectivity
6
2125-2129
Reflectivity aloft still increasing
Inbounds at 1.5 kft now 55 kts
Storm no longer a threat
W 7
2130-2134
Inbounds increase in speed and area
Reflectivity aloft still increasing
BWER
Storm top divergence >100 kts
Intense winds becoming concentrated along the highway
8
2135-2138
Low-level winds have ramped up
BWER
Cloud top registers at 55 kft
Evolution in velocity data suggests storm isn't going to weaken
End

Figure 4.5. A timeline of key observations made in the reflectivity (orange) and velocity (green) AOIs with respect to the northern storm (solid box) and southern storm (dashed box). Time is provided in the arrow for each period (top row) with corresponding case time (UTC) (bottom row). The timings of decisions to issue a warning are indicated by a red “w.”

4.5 Future Applications

The short study presented in this article demonstrates how a forecaster's eye gaze data can be used to understand in greater detail where a forecaster's attention is pointed to and how their attention changed with time. In this instance, we found that the forecaster's fixations changed as a function of the stimulus. We were able to capture his different styles of interrogation of reflectivity and velocity data, and understand how the changing weather scenario impacted the counts and durations of his fixations. Important to our interpretation of trends observed in the fixation measures was the retrospective recall. Together, the eye gaze data and retrospective recall quantified and contextualized the forecaster's cognitive processes, providing a full picture of what, how, and why he was looking at certain points on the screen. The importance of collecting qualitative data to answering the "why" question remains.

The "what" and "how" questions associated with forecasters' decision processes can be answered with more exactness and certainty through eye-tracking. Using eye-tracking to obtain this more informed knowledge about forecaster decision processes may be useful in a variety of applications within operational meteorology. This informed knowledge will become especially important as efforts to become a Weather Ready Nation continue. For example, FACETS is a concept designed to reinvent the watch and warning paradigm from a traditionally deterministic system to one that provides a continuum of probabilistic hazard information (Rothfusz et al. 2014). This change in the watch and warning paradigm requires the development and testing of new tools that will meet forecaster needs (Karstens et al. 2015). The widespread application of eye-tracking methods in usability studies (Jacob and Karn 2003) suggests that eye-tracking will be useful for learning about forecaster-computer interactions and for successfully designing suitable tools.

Eye-tracking may also help determine differences in experienced and expert forecasters' data interrogation strategies and cognitive processes to those of the less-experienced forecaster. Understanding these differences would help in the design of effective training for intern and journeymen forecasters. Furthermore, using eye-tracking to develop a deeper understanding of forecasters' cognitive processes would be helpful in determining whether new types of data and products support or hinder their warning decision processes. For example, the impact of higher-temporal resolution radar data on forecasters' warning decision processes has been studied in the Hazardous Weather Testbed (e.g., Heinselman et al. 2015 and Bowden et al. 2015). Recently, eye-tracking was used in the 2015 PARISE to understand better what these impacts are on forecasters' cognitive processes and their related warning decisions. We expect that collecting forecasters' eye gaze data in addition to their retrospective recalls will better inform us on the specifics of how rapidly-updating radar data affects their data interrogation strategies. For example, we will be able to compare trends in fixation measures between forecasters using radar data of differing temporal resolution, analyze their visual scanning patterns, and develop a more complete picture of their decision processes from start to finish. Finally, introducing eye-tracking research methods to operational meteorology studies provides an opportunity for mutual interdisciplinary knowledge growth between the human factor and meteorology research fields, which can only push the boundaries of our current knowledge.

Chapter 5

Comparing Forecaster Eye Movement Behavior during the Warning Decision Process

Taken in full from: Wilson, K. A., P. L. Heinselman, and Z. Kang, 2016: Comparing forecaster eye movement behavior during the warning decision process. *Submitted to Wea. Forecasting, August 2017.*

Abstract

An eye-tracking experiment was conducted to objectively observe how National Weather Service forecasters distribute their attention and interact with a radar display and warning interface during use of 1-min (experimental group) and 5-min (control group) PAR updates. In addition to demonstrating a new research method for addressing operationally-focused research questions, this experiment was specifically interested in whether forecasters' eye movement behavior can provide further insight into how rapidly-updating radar data impacts the warning decision process. Differences in forecasters' eye movements were therefore analyzed with respect to fixation measures (i.e., count and duration) and visual scanpath dimensions (i.e., vector, direction, length, position, and duration). These analyses were completed for four stages of the warning decision process: the first five minutes of the case, two minutes prior to warning decisions, the warning issuance process, and updates to warnings. While the control and experimental groups' fixation measures were generally similar throughout the four stages, comparisons of the scanpath dimensions detected differences in forecasters' eye movements. Video footage

and retrospective recall data were examined to illustrate how forecasters' interactions with the radar display and warning interface, encounters with technological challenges, and varying approaches to similar tasks resulted in a group's statistically significant (*p*-value<0.05) lower scanpath similarity scores compared to the other group. The findings of this study support the use of eye-tracking research methods for detecting individual differences in forecasters' distributions of visual attention. These individual differences can then be used to better understand why variations in forecasters' warning decision processes occur.

5.1 Introduction

Understanding the forecaster warning decision process is a complex task that has been at the forefront of PARISE since 2010. Learning about potential impacts of rapidly-updating PAR data on forecasters' warning decision processes requires not only an assessment of performance, but an in-depth analysis of how forecasters acquire, make sense of, and use information to provide the best possible warnings (Heinselman et al. 2012; Heinselman et al. 2015; Bowden et al. 2015; Bowden and Heinselman 2016). Other studies within the NOAA Hazardous Weather Testbed have evaluated forecasters' use of the Geostationary Operational Environmental Satellite R (GOES-R) series observing capabilities (Goodman et al. 2012), real-time numerical model analyses (Smith et al. 2014; Calhoun et al. 2014), a probabilistic hazard information tool (Karstens et al. 2015), and newly developed Multi-Radar Multi-Sensor products (Smith et al. 2016). To carry out these evaluations, qualitative methods including observations, surveys, discussions, interviews, and blog posts have been used. Furthermore, in PARISE, cognitive task analysis methods have been applied to obtain detailed insight into what forecasters see, think, and do when presented with radar data of different update speeds (e.g., Heinselman et al. 2015; Bowden and Heinselman 2016). Referred to as the Recent Case Walkthrough (Hoffman 2005), this method requires forecasters to retrospectively recall their thought processes as they watch a playback video of their onscreen activity that was recorded during simulated warning operations. Additionally, as forecasters recall their thought processes step-by-step, they are asked probing questions that tend to focus on times when warning decisions were made.

Much has been learned from retrospective recall data about how faster radar updates can impact forecasters' warning decision processes during different types of severe weather scenarios. However, these data have also brought to light how complex forecasters' warning decision processes can be, and that the use of qualitative methods alone is an insufficient approach for obtaining detailed observations and a comprehensive understanding of forecasters' cognition. Therefore, a more objective method was sought that could both better capture the intricate activity occurring within a forecaster's mind, and address some of the limitations inherent in qualitative methods (i.e., accuracy and completeness of retrospective recall data).

Research studies have shown that our attention is primarily directed to what we are looking at. Also known as the eye-mind hypothesis (Just and Carpenter 1976b; 1980), this connection between our thoughts and our eye movements means that eye tracking can be used to better understand what is happening inside a forecaster's mind when they are presented with radar data. The use of eye-tracking methods was first applied in reading studies (Rayner 1998; Duchowski 2002; Henderson and Ferreira 2004). These initial studies identified two types of eye movement behavior: fixations and saccades.

Fixations describe times when the eye is relatively still and saccades describe the very fast eye movements that occur between fixations. Saccadic suppression effects mean that information is only acquired and processed during fixations (Henderson and Ferreira 2004). Given that reading studies found fixation and saccadic activity to depend on the text that is being read, eye tracking was identified as a useful method for learning about how language is processed.

Applications of eye tracking to study other human cognition was also demonstrated in free-viewing tasks, in which static images rather than text were presented as the stimulus. Early studies used free-viewing tasks to prove that the location of fixations was not random. Rather, fixations occurred more frequently in the most semantically and visually rich regions of an image (e.g., Buswell 1935; Yarbus 1967). This observation was important because it provided evidence that visual processing behavior, as observed through eye movements, is an important representation for attention. More recently, eye tracking has been used in a variety of visual search tasks. For example, research studies focused on web design and marketing have learned much about how the general population attend to and gather information from computer displays, advertisements, and package designs (e.g., Djamasbi et al. 2010; Hervet et al. 2011; Clement et al. 2013; Gidlöf et al. 2013; Qang et al. 2014). Additionally, eye tracking has been used to better understand the visual and cognitive processes of professionals that make life-saving decisions. Within the medical field, many studies have examined the visual search behavior of radiologists tasked with detecting abnormalities and diagnosing medical conditions (e.g., Wood et al. 2013; Manning et al. 2014; Giovinco et al. 2015; Bertram et al. 2016). A review of decision-making research within the medical field found that eye tracking was most frequently used in medical imaging studies, since it is more practical to collect eye movement data during inspection of static scenes compared to within dynamic settings such as an operating theatre (Al-Moteri et al. 2017). In aviation research, eye tracking has been used to study they eye movements of pilots in the cockpit and air traffic controllers on the ground (e.g., Hauland 2008; Sullivan et al. 2011; Van de Merwe et al. 2012; Kang and Landry 2014, 2015; Yu et al. 2016). A common interest in these medical and aviation studies is how visual scanning patterns compare between novice and expert professionals, and whether observed differences can inform training material to improve performance.

Despite the growing popularity of eye-tracking methods in other research domains, eye movement data has been collected in only a handful of meteorology studies. Drost et al. (2015) used eye movement data to analyze what impact a weathercaster's gesturing would have on viewers during a televised weather forecast, and found that while the gesturing influenced where viewers looked, it did not affect what they remembered. Eye tracking was also used to assess the impact of legend color and content on participants' abilities to correctly interpret hurricane storm surge graphical information (Sherman-Morris et al. 2015). While statistically significant differences were not found in performance for use of legends differing in color and content, participants' eye movement data indicated that they struggled most when legends were presented in shades of blue and with values in feet (Sherman-Morris et al. 2015). In an exploratory sense, Wilson et al. (2016) assessed the feasibility of eye tracking as a research method for improving and building upon the current understanding of forecasters' warning decision processes. Without previous examples of NWS forecasters' eye movement data, a simple question was whether such eye movement data would make sense and be representative of a forecaster's experienced cognitive activity. In this short study, Wilson et al. (2016) collected a single NWS forecaster's eye movement data as they interrogated radar data during simulated warning operations. This participant's retrospective recall was also collected following the simulated event, just as in previous PARISE studies (Heinselman et al. 2015; Bowden and Heinselman 2016). Comparing trends in these eye movement data to the participant's retrospective recall, this study concluded that the eye movement data were able to successfully capture important events during the simulation (e.g., change in expected threat and the subsequent redistribution of attention), and were therefore representative of the forecaster's warning decision process (Wilson et al. 2016).

The findings from Wilson et al.'s (2016) study supported the use of eye tracking as a method for observing the visual attention of a forecaster in an objective manner, in real time, and with greater temporal detail and accuracy than what has been observed before. These findings motivated a larger-scale study that we present in this paper. Of particular interest is how forecasters' eye movement behavior compare with respect to 1) fixation measures and 2) overall visual scanning patterns during use of different radar update speeds. Given that previous studies have shown that the use of 1-min, 2-min, and 5-min radar update speeds can impact performance and overall situational awareness (Heinselman et al. 2015; Bowden et al. 2015; Bowden and Heinselman 2016; Wilson et al. 2017), we were specifically interested in whether differences in forecasters' related warning decision processes would be evident in their eye movement behavior. This study explores what, if any, differences existed between forecasters' eye movements while they worked a single weather event using 1-min or 5-min PAR updates. Retrospective recall and video data are used to understand these differences in the broader context of the warning decision process. Moreover, the findings from this research contribute to our current limited knowledge of how eye tracking can be applied to address operational meteorology research questions and what forecasters' eye movement data can teach us about the human component of weather forecasting.

5.2 Methodology

5.2.1 Experimental Design

Over six weeks in the summer of 2015, 30 NWS forecasters from 25 WFOs visited the NOAA Hazardous Weather Testbed in Norman, Oklahoma to participate in the 2015 PARISE (Wilson et al. 2017). The largest of its kind, this most recent PARISE was comprised of three studies: the traditional experiment (Wilson et al. 2017), the eye-tracking experiment, and the focus group. This paper presents results related to the eye-tracking experiment only. In this eye-tracking experiment, forecasters worked a one-hour long event independently in simulated real time. Forecasters were randomly assigned to either a control or an experimental group, which determined whether they were presented with 5-min or 1-min PAR updates, respectively. Both groups had an equal number of participants. During the case, forecasters were provided with reflectivity and velocity base products only, and were able to display these data using the WDSS-II software (Lakshmanan et al. 2007). Given that not all forecasters were familiar with WDSS-II, training on how to setup and navigate through the radar data and issue warning products was provided. A warning generation (WarnGen) tool similar to what forecasters use in operations was developed for WDSS-II, and all issued warning products were recorded in an electronic database. As in previous PARISE studies, a pre-briefing video lasting several minutes was provided prior to working the case to allow forecasters to form expectations for how the weather event may unfold. This video described the environmental conditions associated with the upcoming weather event, and showed prior radar and satellite data leading up to the case start time. Once forecasters had watched the pre-briefing video, they were asked to work the weather event with their normal approach and to make warning decisions if considered necessary.

5.2.2 Weather Scenario

The chosen weather scenario included a multicell severe hail and wind event that occurred during 2230–2330 UTC 8 July 2014. In addition to meeting a suitability criteria for experimental testing (i.e., uninterrupted radar observations for a sufficient duration), discussions with the NWS forecaster that worked the event in real time influenced the case selection. After viewing 1-min PAR updates of this event, the forecaster reported being able to better track cycling trends in rapid core development aloft compared to when he had used the 5.1 min WSR-88D volume updates (personal communication, Charles Kuster). We therefore anticipated that this case would present forecasters with an opportunity to demonstrate differences in their warning decision processes when using 1-min or 5-min PAR volumetric updates.

In this scenario, the 90° PAR sector scanned towards the southeast and encompassed two areas of storms (Fig. 5.1). The storm in the western portion of the sector is referred to as the McClain storm, and the storm in the eastern portion of the sector is referred to as the Pontotoc storm. The discreet nature of these storms further encouraged the selection of this case, since it allowed for a clear-cut analysis of how attention was distributed between the storms. According to the official NWS Storm Data records (https://verification.nws.noaa.gov), only the McClain storm was associated with severe hail (at 2304 UTC and 2328 UTC) and wind (at 2325 UTC) reports. Although the Pontotoc storm was not associated with severe weather reports in Storm Data, this storm presented more impressive characteristics in radar data and had higher values of Maximum Estimated Size of Hail (Witt et al. 1998) than the McClain storm. Therefore, it is possible that the Pontotoc storm also produced severe weather, but that it was not observed nor reported.

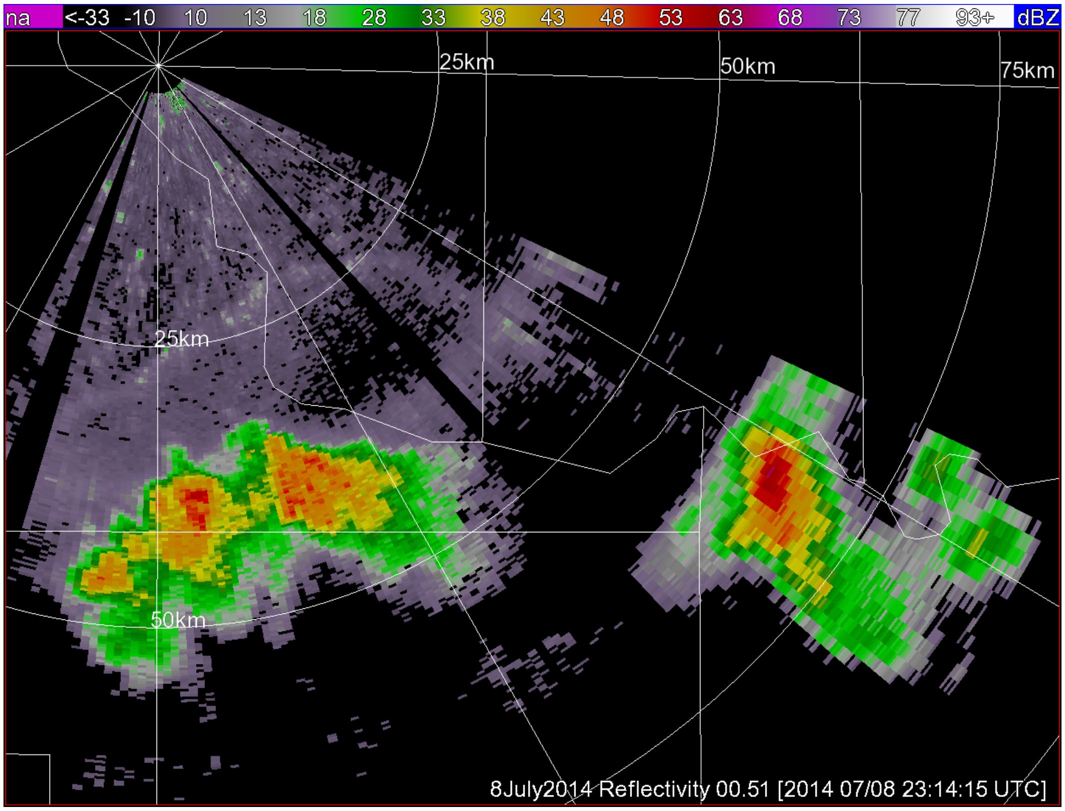


Figure 5.1. Snapshot of the 0.5° reflectivity data at 2314 UTC 8 July 2014.

5.2.3 Data Collection

The Tobii TX300 eye-tracking system was used to collect forecasters' eye gaze data. This remote video-based system uses infrared illumination to track pupil and corneal reflection. More specifically, dark-pupil eye-tracking methods were used, such that the infrared illumination was positioned away from the optical axis, causing the pupil to appear darker than the iris. The video camera in the eye-tracking system acquired an image of the eye at a sampling rate of 300 Hz. Through the use of image processing algorithms, the dark pupil and corneal reflection were identified, and geometrical calculations, as well as information from each forecaster's calibration, were used to map the point of vision to *x* and *y* coordinates on the computer screen.

The calibration procedure each forecaster completed prior to beginning the case required them to watch the computer screen and follow a series of dots as they appeared. To ensure calibration was completed successfully, we also asked each forecaster to spend a short time browsing a webpage. We used this sample of eye gaze data to ensure that the eye-tracker captured their point of vision accurately. Once calibration was completed, the Tobii TX300 was used to collect each forecaster's eye gaze data for the full duration of the weather scenario. The remote eye-tracking system was positioned beneath the computer screen, and although forecasters had to remain relatively still while working the case, some gentle head movements were allowed.

At the end of the case, the collected eye gaze data was checked to ensure that the gaze sample was sufficient. The gaze sample is a measure that indicates the proportion of samples that were collected successfully, is given as a percentage, and is considered acceptable for values of at least 75% (Hvelplund 2014). Data loss resulting in gaze samples below this value can occur due to difficulty in detecting the pupil and corneal reflection, possibly due to a person's eye color, eye shape, use of eyewear, or use of makeup. Furthermore, visual inspection of the overlaid eye gaze data on the screen recording was important for ensuring sufficient accuracy and precision of forecasters' eye gaze data. Based on these data quality checks, six data sets were removed from the analysis, and the results presented in this paper are therefore based on eye gaze data belonging to twelve participants in each group.

Each forecaster also provided a retrospective recall of their warning decision process using the Recent Case Walkthrough method. As described in the introduction, this method has also been used extensively in the 2012 and 2013 PARISE studies (Heinselman et al. 2015; Bowden et al. 2016). We asked forecasters to verbalize their thought processes while watching a playback video of their onscreen activity. Concurrently, the assisting researcher typed these verbalizations into a timeline. Probing questions were used to gather further insight into why forecasters made warning decisions.

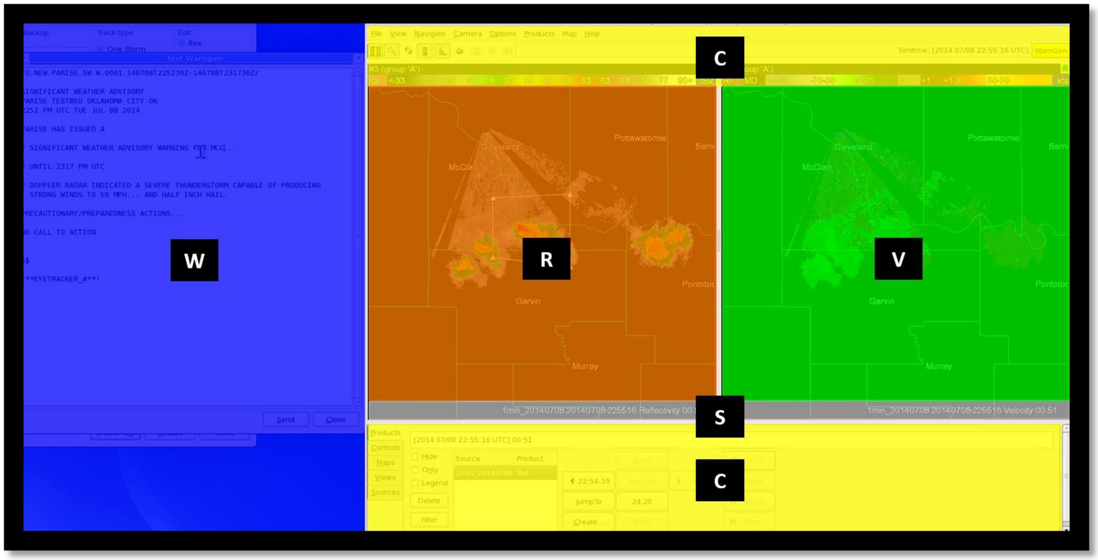
5.2.3 Data Analysis

5.2.3.1 Fixation Identification

Fixation events are of most interest because it is during these times when humans process information (Henderson and Ferreira 2004). To identify fixation events, the raw eye gaze data was parsed through a velocity-threshold identification (I-VT) algorithm using the Tobii Studio 3.3.0 software (Komogortsev et al. 2010; Olsen 2012; Tobii 2017). This algorithm's output lists the timestamp, duration, and *x* and *y* position for each fixation. The *x* and *y* positions are based on a pixel grid system of the computer screen (1920 pixels by 1080 pixels). Eye gaze velocity is described in terms of visual angle ($°s^{-1}$) and is calculated as the angle between two samples divided by their separation in time. To reduce measurement noise effects, angular velocity is calculated for a 20*ms* window which is centered on the sample of interest (Olsen 2012). The timestamp and position information of the first and last sample of the window determine the angular velocity of the center sample. Samples having an angular velocity below the default velocity threshold parameter ($30°s^{-1}$) are classified as fixations (Olsen 2012; Bojko 2013). Adjacent fixations may either remain separate or be merged into a single longer fixation depending on the time and visual angle between them. The "max time between fixations" parameter is given as 75*ms* (allowing for blink events), and the "max angle between fixations" parameter is set at 0.5° (Komogortsev et al. 2010). Two adjacent fixations become merged if the time and angle between them is less than or equal to these parameter values. Finally, a minimum fixation duration parameter of 60*ms* was chosen. This minimum fixation duration was chosen because fixations during reading studies have shown to last between 60*ms* –500*ms* (Liversedge and Findlay 2000). All fixations with durations shorter than 60*ms* were discarded.

5.2.3.2 Areas of Interest and Fixation Measures

In addition to identifying eye fixation events, the Tobii Studio 3.3.0 was used to manually draw AOIs that define separate spaces on the computer display (Holmqvist et al. 2011; Bojko 2013). These AOIs represent different semantic content, including: reflectivity data, velocity data, control icons, radar scan information, and the WarnGen interface (Fig. 5.2). The two control icon areas were combined in the analysis. All identified fixations were tagged with the AOI in which they occurred. The AOI-based labelling of fixations is useful for comparing forecasters' visual processing behaviors within these spaces for different portions of the warning decision process. While many different types of fixation measures exist, two of the most commonly used measures are count and duration (Jacob and Karn 2003). We can assess within each AOI how many times forecasters fixated (count) and on average how long those fixations lasted (duration). Higher fixation counts within an AOI indicate that the information was more noticeable or important to the participant, while an AOI associated with longer fixation durations indicates that the information was either more difficult to extract or more engaging to the participant (Poole and Ball 2006; Bojko 2013).



```
/O.NEW.PARISE.SW.W.0001.140708T225230Z-140708T231730Z/

SIGNIFICANT WEATHER ADVISORY
PARISE TESTBED OKLAHOMA CITY OK
2252 PM UTC TUE JUL 08 2014

PARISE HAS ISSUED A

* SIGNIFICANT WEATHER ADVISORY WARNING FOR MCC...

* UNTIL 2317 PM UTC

* DOPPLER RADAR INDICATED A SEVERE THUNDERSTORM CAPABLE OF PRODUCING
  STRONG WINDS TO 50 MPH... AND HALF INCH HAIL.

PRECAUTIONARY/PREPAREDNESS ACTIONS...

NO CALL TO ACTION


$$

**EYETRACKER_#**!
```

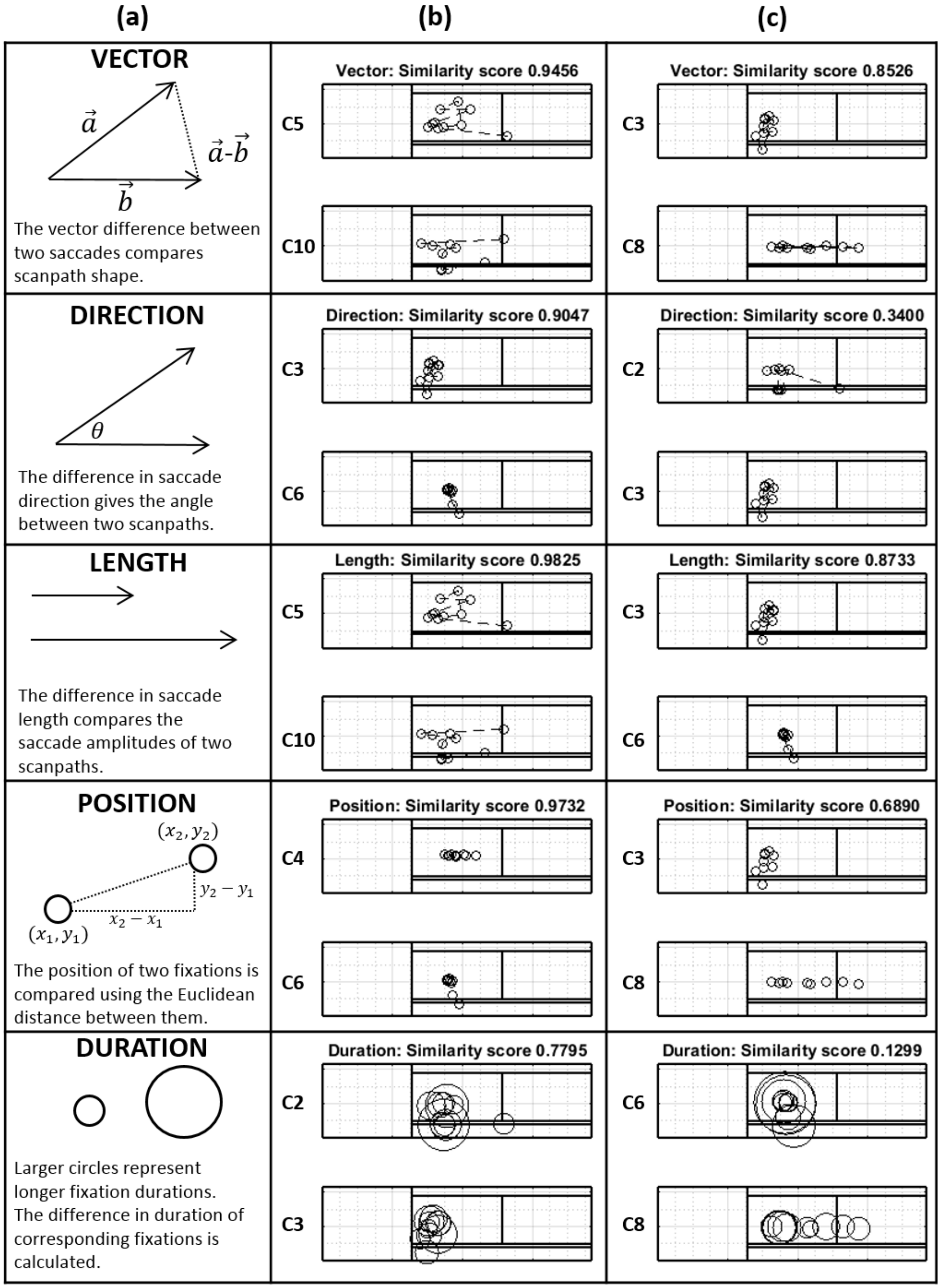
Figure 5.2. Areas of interest are identified for the reflectivity data ("R", orange), velocity data ("V", green), control icons ("C", yellow), radar scan information ("S", grey), and the WarnGen interface ("W", blue). Note that the WarnGen interface appeared only when the forecaster selected to use it.

5.2.3.3 Scanpath Comparisons

Fixation measures are useful for obtaining an overall impression of how visual attention is distributed across AOIs for a given timeframe. These measures can be used to indicate whether the control and experimental groups visually attended to the different AOIs in a similar manner or not. This type of analysis was useful during Wilson et al.'s (2016) initial eye-tracking study, where differences in the participating forecaster's fixation measures across the Reflectivity and Velocity AOIs corresponded to an anticipated change in the weather threat and alteration of attention resources accordingly. However, these bulk measures are not good at representing how attention is distributed over time. Additionally, the spatial resolution of fixations is reduced to the size of the AOIs, meaning that the spatial distribution of fixations within an AOI is not represented either. How fixation behavior changes in time and space is an important consideration if forecasters' underlying cognitive processes during this simulation are to be understood. Therefore, in addition to average AOI fixation measures, the sequence of fixations in time and space is examined with AOI boundaries removed.

Noton and Stark (1971) first described these sequences in an abstract sense as the viewing patterns of a person, and termed this idea a "scanpath." Today, the term scanpath is given the physical definition of "the route of oculomotor events through space within a certain timespan" (Holmqvist et al. 2011). Early applications of scanpath analysis required visual inspection of the temporal and spatial ordering of fixations. However, analysis methods have since developed and there is now a variety of ways to compare and quantify scanpath data (e.g., Anderson et al. 2015). Comparisons of scanpaths become especially important when trying to understand similarities or differences in visual processing behaviors of multiple people or of the same person but at different times. The variety of comparison methods differ in how they treat a sequence of fixations and what aspects of the scanpaths they are able to measure. In our study, it is essential to maintain temporal ordering of the entire sequence of fixations. Of the many scanpath comparison methods described in Anderson et al. (2015), only three met this requirement: String Edit Distance, ScanMatch, and MultiMatch. Still, both String Edit Distance (Levenshtein 1966) and ScanMatch (Cristino et al. 2010) rely on AOI-base methods, meaning that spatial resolution of fixation position is lost. This reduction in spatial information means that the shape of scanpaths cannot be represented adequately in similarity calculations (Jarodzka et al. 2010).

Acknowledging this limitation, Jarodzka et al. (2010) developed a new scanpath comparison method called MultiMatch. This method is based on vector representations of scanpaths (i.e., in *x* and *y* space) and preserves a number of aspects, including: the position and duration of fixations, the shape of scanpaths, and the length and direction of scanpath saccades. The MultiMatch scanpath comparison method first simplifies participants' scanpaths using amplitude- and direction-based clustering, causing clustering of very short vectors within the same local space and of consecutive vectors with very similar direction (Jarodzka et al. 2010; Dewhurst et al. 2012). Following the simplification of two scanpaths, approximate temporal alignment of vector saccades and fixations is determined using vector shape information and the Dijkstra (1959) algorithm. A more detailed description of this alignment method is given in Jarodzka et al. (2010) and Dewhurst et al. (2012). Once scanpath alignment has been determined, five similarity measures are computed for the paired fixation and saccade vectors of two given scanpaths, and these measures are then averaged to give five similarity scores. These five MultiMatch measures compare the vector, length, direction, position, and duration of two scanpaths (Fig. 5.3). Since the similarity score is calculated differently for each of the five measures, absolute score values cannot be compared across measures. However, the distributions of these similarity scores within the same measure for the control and experimental groups will indicate whether one group has more variable scanpath behavior than the other. Video and retrospective recall data will provide context and explanation for the observed results.


(a)
(b)
(c)
VECTOR
$\vec{a}$
$\vec{a}$-$\vec{b}$
$\vec{b}$
Vector: Similarity score 0.9456
C5
C10
Vector: Similarity score 0.8526
C3
C8
DIRECTION
θ
Direction: Similarity score 0.9047
C3
C6
Direction: Similarity score 0.3400
C2
C3
LENGTH
Length: Similarity score 0.9825
C5
C10
Length: Similarity score 0.8733
C3
C6
POSITION
(x_2, y_2)
$y_2 - y_1$
$x_2 - x_1$
(x_1, y_1)
Position: Similarity score 0.9732
C4
C6
Position: Similarity score 0.6890
C3
C8
DURATION
Duration: Similarity score 0.7795
C2
C3
Duration: Similarity score 0.1299
C6
C8

The vector difference between two saccades compares scanpath shape.

The difference in saccade direction gives the angle between two scanpaths.

The difference in saccade length compares the saccade amplitudes of two scanpaths.

The position of two fixations is compared using the Euclidean distance between them.

Larger circles represent longer fixation durations. The difference in duration of corresponding fixations is calculated.

Figure 5.3. The a) five MultiMatch measures with corresponding examples of scanpaths that have b) relatively higher similarity scores and c) relatively lower similarity scores for two control participants. Adapted from Dewhurst et al. (2012).

5.2.3.4 Defined Stages

Previous PARISE studies have observed that when working weather events in simulated real time, there are clear stages in the warning decision process that are common among all forecasters. To better understand similarities and differences in forecasters' cognitive processes during times in which they are engaged in the same task, we chose to focus our analysis of the eye movement data on four stages: 1) the first five minutes of the case, 2) two minutes prior to warning decisions, 3) the warning issuance process, and 4) the first update on the McClain and Pontotoc storms. The timing of these stages for all 24 participants was identified using video and retrospective recall data (Fig. 5.4), and their corresponding eye movement data was extracted for analysis. For each of these stages, participants' fixation count and mean fixation duration were calculated, and the five MultiMatch measures were computed for all possible participant scanpath combinations within each group.

In this study, all forecasters issued a severe thunderstorm warning at least once on the McClain storm and once on the Pontotoc storm (Fig. 5.5). Eleven control and ten experimental participants also issued a second severe thunderstorm warning on the McClain storm (Fig. 5.5). Given that these were major warning decisions across both groups, the warning issuance process for each of these three decisions is included in the analysis. Finally, updates to these warnings were completed through the issuance of severe weather statements (SVSs). Some forecasters issued many more SVSs than others, but eleven experimental and all control participants issued at least one SVS on the McClain storm, and six participants in each group issued at least one SVS on the Pontotoc storm (Fig. 5.5). For the fourth stage, we therefore focus on the first SVS issuance for each of these storms.

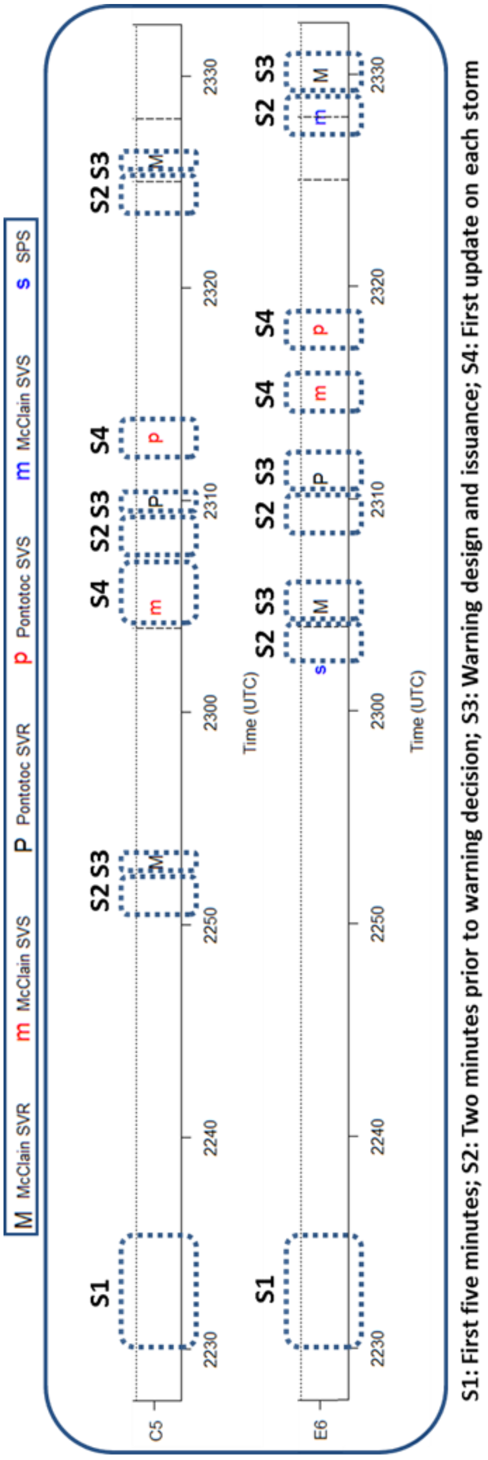


Figure 5.4. Examples of the four defined stages occurring within control participant C5's and experimental participant E6's warning decision process. With the exception of the first five minutes, the general duration of each stage lasted 1–3 minutes. Severe thunderstorm (SVR) issuance times are shown for the McClain storm (M in black) and Pontotoc storm (P in black), and first SVS issuance times for the McClain storm (m in red) and Pontotoc storm (p in red). Additional issuance decisions potentially impacting analyzed stages include issuance of a special weather statement (s in blue) and issuance of a second SVS for the McClain storm (m in blue). Vertical dashed lines indicate the timing of severe weather reports (see Fig. 5.5).

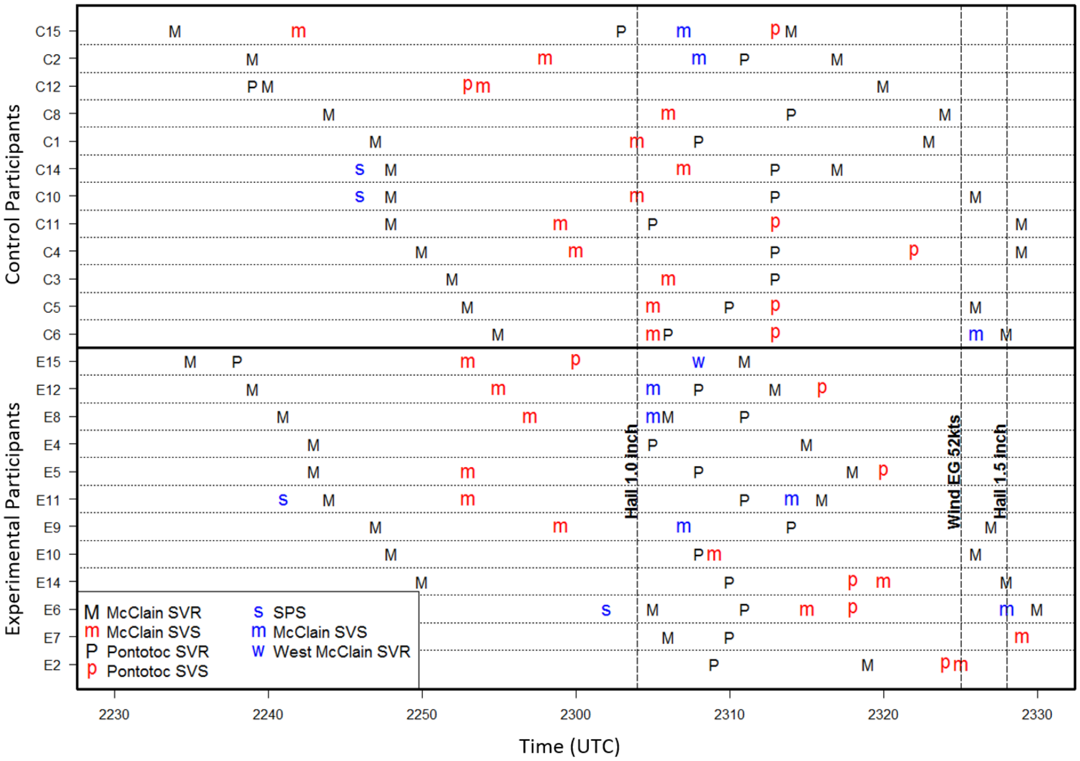


Figure 5.5. Control (top) and experimental (bottom) participants' warning products issued during the weather scenario. Markers are the same as in Figure 5.4. Vertical dashed lines indicate the timing of the severe weather reports associated with the McClain storm.

5.3 Results

5.3.1 First Five Minutes

The first five minutes characterizes a time in which forecasters were busy loading their radar data and familiarizing themselves with the weather scenario. Video and retrospective recall data show that forecasters in both groups spent much of their time sampling the reflectivity profiles of the McClain and Pontotoc storms, frequently moving back and forth between the two storms while climbing in elevation for vertical comparison. The eye fixation measures of the control (5-min PAR updates) and experimental (1-min PAR updates) groups reflect this observed behavior. Attention was given primarily to the Reflectivity AOI, with the median fixation count in this AOI exceeding that of any other AOI ($Count_{con}(SD) = 404\ (34)$ and $Count_{exp}(SD) = 367\ (77)$). The second highest median fixation count for both groups occurred within the Velocity AOI ($Count_{con}(SD) = 111\ (69)$ and $Count_{exp}(SD) = 94\ (53)$). Only the deeper McClain storm was visible at higher elevations, and for most forecasters a choice was made to prioritize attention on this storm. These observations led one participant in each group to issue a severe thunderstorm warning on the McClain storm (C15 and E15), while one additional experimental participant (E5) also prepared a similar warning (Fig. 5.5). All other forecasters, however, did not visit the WarnGen AOI during this time.

Both groups' scanpaths were relatively more similar during these first five minutes compared to the later defined stages (Fig. 5.6). Differences in the groups' similarity scores for four of the five MultiMatch dimensions were not statistically significant, indicating a comparable level of variability in forecasters' scanpath behavior within each group. However, the groups did differ with respect to fixation duration (*p-value*<0.001), with the experimental group's lower similarity scores indicating more differences in their processing of these data (Fig. 5.6e). The experimental group's larger variation in fixation duration was most evident within the Reflectivity and especially Velocity ($Dur_{con}(SD) = 397ms\ (48\ ms)$ and $Dur_{exp}(SD) = 443ms\ (132\ ms)$) AOIs, where the experimental group's spread in fixation duration is notably greater than the control group's.

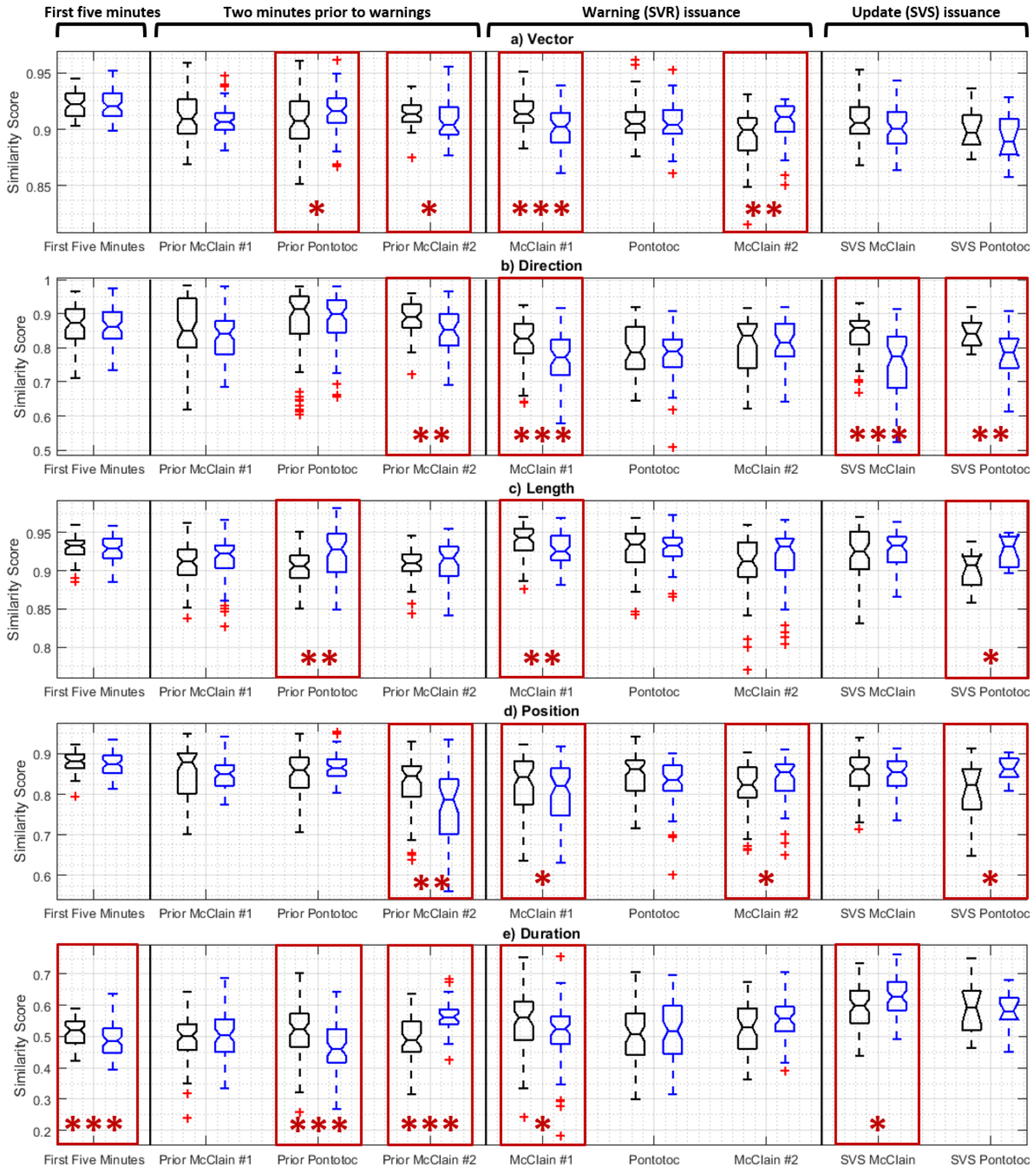


Figure 5.6. Boxplot distributions of similarity scores for the five MultiMatch measures a) vector, b) direction, c) length, d) position, and e) duration for the control group (left position, black) and experimental group (right position, blue). Red boxes indicate distributions that are significantly different according to the Wilcoxon-Mann-Whitney rank-sum test (*p-value<0.05, **p-value<0.01, and ***p-value<0.001). Red crosses (+) indicate outlier values that are less (greater) than 1.5 times the lower (upper) quartile.

5.3.2 Two Minutes Prior to Warning Decision

Forecasters' eye movement behavior in the two minutes proceeding a warning decision were analyzed for three occasions. No differences in fixation measures or any of the five MultiMatch similarity scores were found to be statistically significant between the groups prior to the first warning on the McClain storm (Fig. 5.6). For most participants, interrogation continued in a manner similar to the first five minutes, such that fixations in the Reflectivity AOI were three to four times as frequent as those in the Velocity AOI. However, forecasters' relative lack of references to the Pontotoc storm in the retrospective recall and video data show that forecasters shifted their attention more so to the McClain storm in the two minutes leading up to their decisions to warn.

Although fixation measures between the control and experimental groups were also not statistically significantly different in the two minutes prior to the Pontotoc warning, greater variability in the control group's scanpath behavior was observed in the vector and length MultiMatch dimensions (Fig. 5.6a, c). The lower vector similarity scores were due to C12's chosen method for navigating through the radar data. While C12 preferred to click on icons located in the Control AOIs (Fig. 5.7a), all other forecasters followed the taught method of toggling with computer keys. C15 was responsible for the lower length similarity scores because of their decision to focus interrogation only on the Reflectivity AOI to "Find hail cores aloft" (Fig. 5.7b). The overall shape of C12's scanpath and the shorter saccades belonging to C15's scanpath prior to the Pontotoc storm warning decision was visibly different than that of C10's. Participant C10's scanpath (Fig. 5.7c) is a more typical representation of how forecasters spent their time prior to the Pontotoc warning decision. As this representative gaze plot shows, although forecasters' attention was distributed heavily within the Reflectivity AOI prior to the Pontotoc storm warning, they also tended to check the Velocity AOI to analyze storm top divergence, midlevel rotation, and low-level wind signatures (Fig. 5.7c).

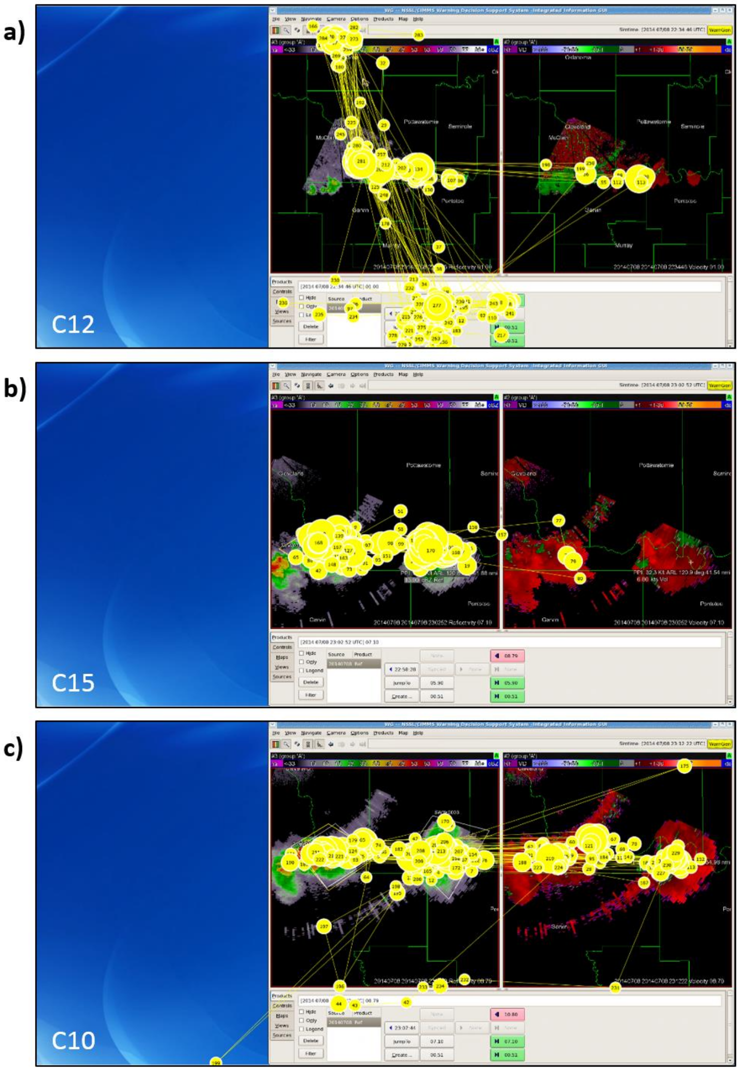


Figure 5.7. Gaze plots depicting the scanpaths of participants a) C12, b) C15, and c) C10 in the two minutes prior to the Pontotoc storm warning decision. Circles represent fixations, the circle center identifies the fixation location, and the circle size characterizes the fixation duration. Lines between fixations represent the corresponding saccades. The background screenshot is the final frame from the period depicted.

While two participants' unusual fixation behavior explained the control group's lower scanpath similarity prior to the Pontotoc warning, more prominent group differences occurred prior to the second McClain warning decision. On average, experimental participants fixated twice as often in the WarnGen AOI than control participants ($Count_{con}(SD) = 16\ (37)$ and $Count_{exp}(SD) = 34\ (81)$), while control participants fixated more frequently within the Velocity AOI ($Count_{con}(SD) = 44\ (31)$ and $Count_{exp}(SD) = 28\ (29)$) but for a statistically significant shorter mean duration ($Dur_{con}(SD) = 367\ ms\ (81\ ms)$ and $Dur_{exp}(SD) = 450\ ms\ (49\ ms)$) (*p*-value= 0.0133). Whereas the higher Velocity AOI fixation count corresponds to control participants' more frequent observations of the McClain storm's strengthening low-level wind signatures, experimental participants' greater use of WarnGen largely explains their statistically significant lower similarity scores for four of the five MultiMatch dimensions (Fig. 5.6). For example, E11's low similarity scores were due to spending much of these two minutes issuing a cancellation on the first McClain warning having previously seen a downward trend in the reflectivity core (Fig. 5.8a). Following this cancellation, he "Noticed a gigantic three body scatter spike coming off that core that had 50dBZ at 32kft," and quickly decided to issue a second warning on this storm. In addition to E11's cancellation, observations of increasing reflectivity values aloft (and an associated updraft pulse) coupled with a storm report prompted E6 (Fig. 5.8b) and E8 to update the first McClain storm warning during these two minutes. E15's use of WarnGen during this time was because of his decision to issue a warning on storm development to the west of the McClain storm given the strengthening 1-min trends in its reflectivity core. Unlike these four experimental participants, others within this group used their time to focus only on the radar data and produced a scanpath evidently different to those that carried out WarnGen-based tasks (e.g., E14, Fig. 5.8c). Although these other experimental participants checked the Pontotoc storm intermittently, most of their time was spent on the McClain storm "because it had various reports and the warning [was] coming close to expiration" (E14).

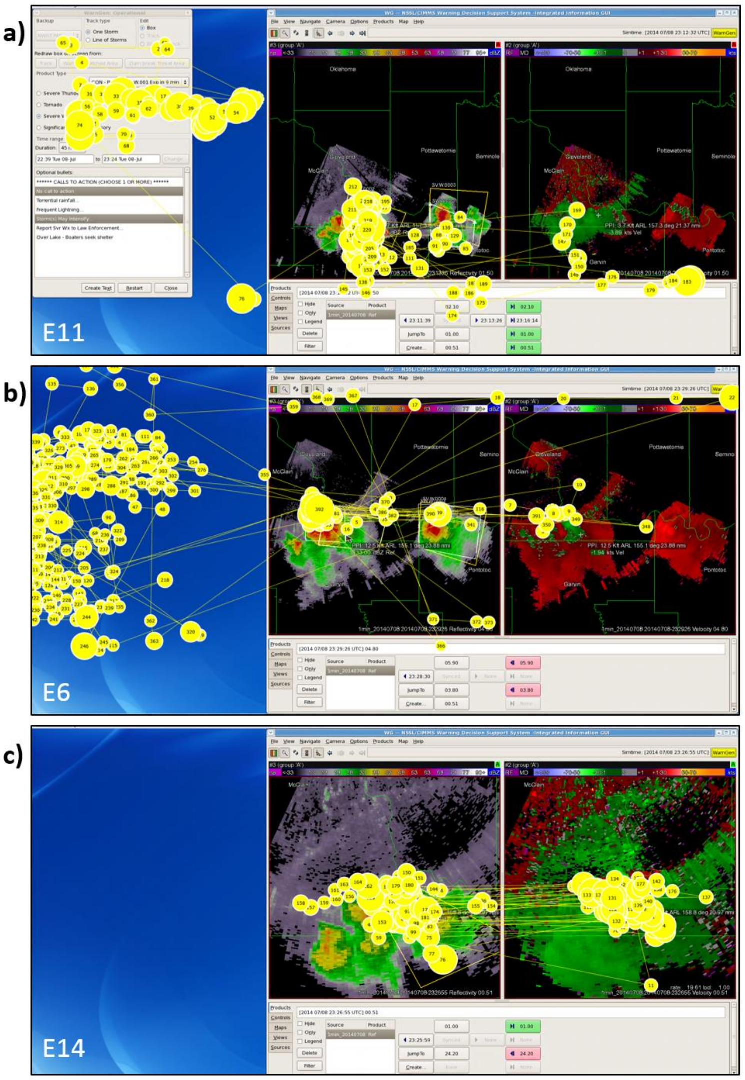


Figure 5.8. Gaze plots depicting the scanpaths of participants a) E11, b) E6, and c) E14 in the two minutes prior to the second McClain storm warning decision. Note that the WarnGen tool could be toggled on and off at any time and sometimes did not appear in the final screen capture.

5.3.3 Warning Issuance Process

The warning issuance process usually took 1–3 minutes to complete, and the video data show that most forecasters followed a typical routine. This routine involved forecasters: loading WarnGen, using the "drag me to storm" icon to set their polygon, adjusting polygon vertices, looping reflectivity data (usually at 0.51°), readjusting vertices to better account for storm development and motion, choosing call to actions, creating and scanning the text, and lastly signing and sending the warning. The majority of forecasters' scanpath patterns were thus mostly confined to the Reflectivity and WarnGen AOIs.

Although forecasters' fixation measures were comparable across both groups during the issuance of the first McClain warning, several participants' deviation from the typical issuance routine resulted in statistically significant lower scanpath similarity scores within the experimental group for all five MultiMatch dimensions (Fig. 5.6). For example, E2 did not feel the urgency to warn given that the "Situation was not rapidly evolving" and he could "Afford to spend time on [the] warning product [to get a] good handle on what's going on" (Fig. 5.9a). While issuing the first McClain storm warning, E2 spent considerably more time than other forecasters watching storm trends, ensuring that the Pontotoc storm did not require his attention, and as he reported, "Nitpicking small details." Similarly, E12 used time while designing the warning to analyze trends in radar data and carefully consider what threats to include in the warning, which call to actions to select, and for how long the warning should be issued (Fig. 5.9b). Additionally, two forecasters struggled with technical disruptions when issuing the warning. E11 struggled to set the polygon correctly because he "[Couldn't] fine tune counties as much as [he would] like," while E14 found that the polygon "Kept snapping around on [him]," causing him to switch between the Reflectivity and WarnGen AOIs frequently and have more broadly distributed fixations across the Reflectivity AOI after repeatedly readjusting the vertices (Fig. 5.9c).

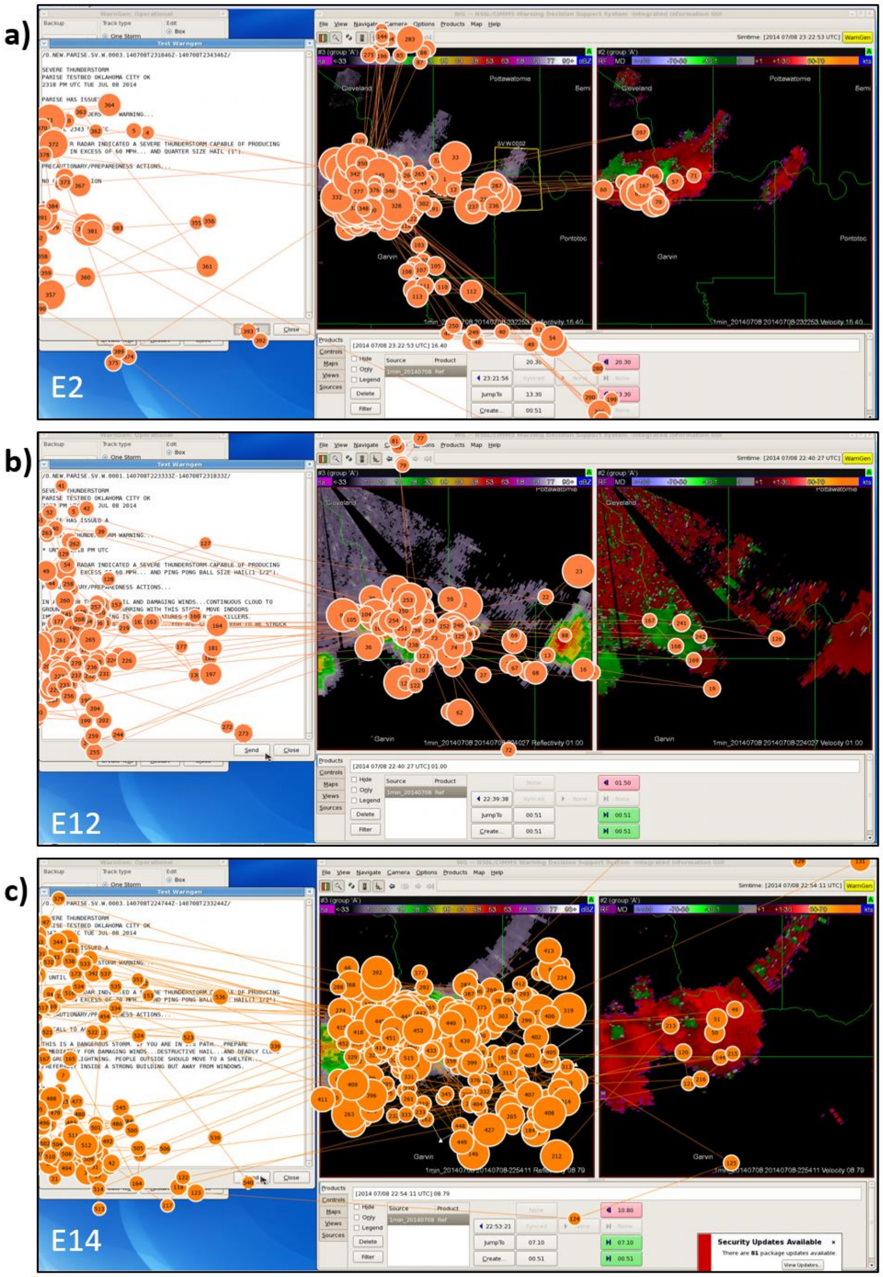


Figure 5.9. Gaze plots depicting the scanpaths of participants a) E2, b) E12, and c) E14 during the issuance of the first McClain storm warning.

The few technical challenges observed during the issuance of the first McClain warning did not arise during the Pontotoc warning issuance and thus did not reduce scanpath similarity among participants. Furthermore, the majority of participants' decisions to issue this warning were prompted within 5–10 minutes of receiving the first hail report, and the timing and reasoning of the Pontotoc storm warning was therefore much more similar than for the first McClain storm warning (Fig. 5.5). It is then unsurprising that forecasters followed the routine warning issuance process for the Pontotoc storm and no statistically significant differences between the control and experimental groups' fixation measures or MultiMatch dimensions were observed (Fig. 5.6).

For most participants, the final warning was issued again on the McClain storm (Fig. 5.5). Unlike the first McClain warning, experimental participants' scanpaths during this second issuance were more similar to one another than the control participants' (Fig. 5.6). The unusual scanpath behavior of three control participants explain why the vector and position similarity scores were statistically significantly lower for this group. First, despite most other participants thinking that the McClain storm continued to pose a severe weather threat, C4 was "Not impressed with the storm" and "reluctantly" decided to issue the second McClain storm warning after receiving all storm reports (Fig. 5.5). He zoomed into the McClain storm during this issuance and transitioned between the Reflectivity and WarnGen AOIs only once (Fig. 5.10a). This single transition is an important aspect of C4's scanpath because it was more typical for forecasters to transition between these two AOIs multiple times during warning issuance. Like C4, C2 also "Did not think the storm was severe enough to warn on again." However, the first hail report associated with the

McClain storm prompted C2 to hesitantly issue a second warning given that his first McClain storm warning was issued early in the case and would soon be expiring (Fig. 5.5). C2’s hesitance was evident in his numerous revisits to the Reflectivity AOI to sample the magnitude of the high-reflectivity core while creating the warning.

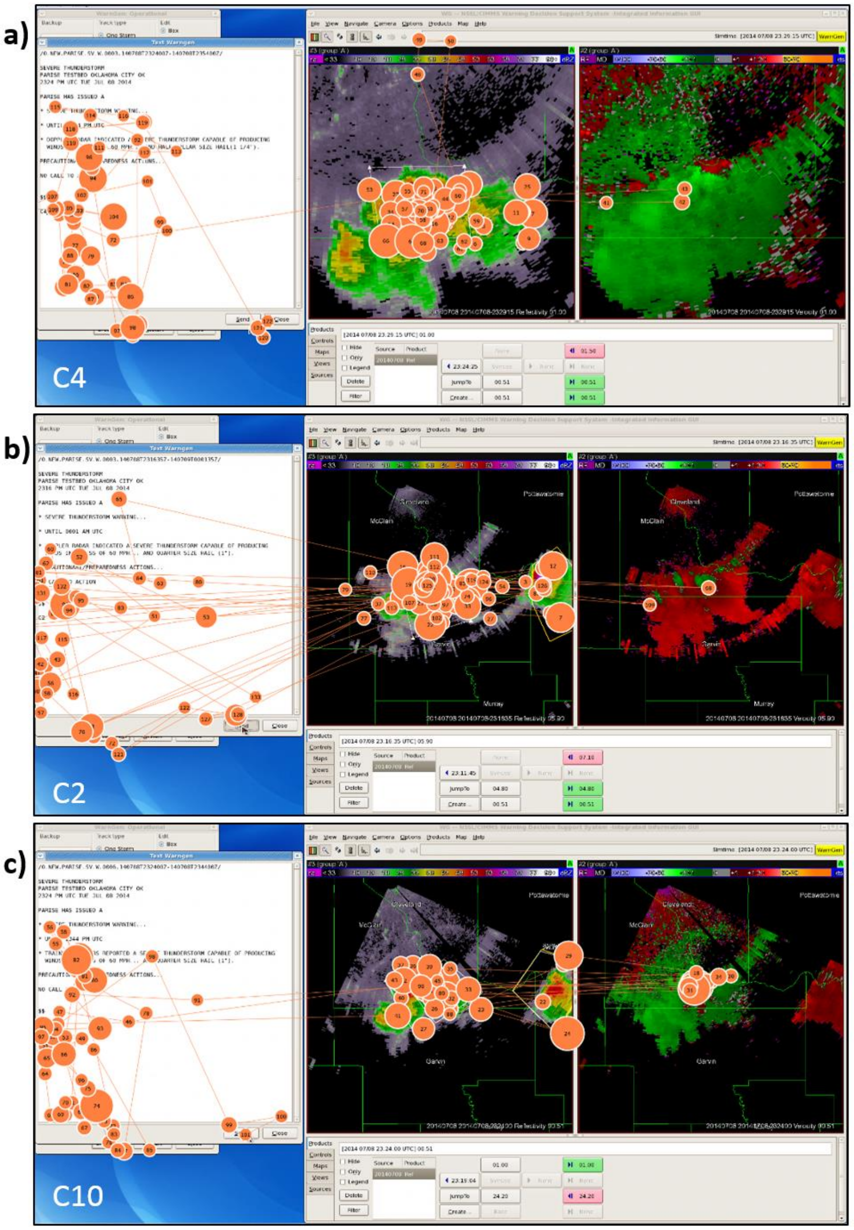


Figure 5.10. Gaze plots depicting the scanpaths of participants a) C4, b) C2, and c) C10 during the issuance of the second McClain storm warning.

This behavior resulted in many more transitions between the Reflectivity and WarnGen AOIs than what was typical of other participants (Fig. 5.10b). The third control participant that presented an unusual scanpath was C10. While the issuance of the second McClain warning was a quick process for this participant, he had previously noted stronger inbound velocities, and therefore visited the Velocity AOI to monitor these data while designing the warning (Fig. 5.10c). If at all, most other participants only glanced in the Velocity AOI during this warning issuance.

5.3.4 Warning Update Process

The timing and reasoning of the first update to the McClain storm warning was more varied among experimental participants than control participants. Whereas a storm report drove more than half of the control participants' decision to issue this SVS, most experimental participants issued this update to provide "maintenance" to the warning by altering the expected weather threat based on radar observations, trimming areas of the warning polygon, or simply providing a continuation of the warning. The experimental group's greater spread in fixation counts within the Reflectivity ($Count_{con}(SD) = 38\ (30)$ and $Count_{exp}(SD) = 17\ (40)$) and WarnGen ($Count_{con}(SD) = 104\ (59)$ and $Count_{exp}(SD) = 109\ (96)$) AOIs, along with their statistically significant lower direction similarity scores (Fig. 5.6), illustrates their more variable scanpath behavior during this warning update compared to the control group. The experimental group's lower direction similarity scores occurred due to participants that either updated the warning with an unusually quick or an unusually extended process. For example, while a couple of experimental participants issued the SVS without changing any aspect of the warning (e.g., E10, Fig. 5.11a), others spent considerable time assessing the radar data, updating the expected weather threat, and carefully adjusting the polygon vertices (e.g., E5, Fig. 5.11b). These contrasting warning update processes were not observed in the control group; rather, all control participants changed at least one aspect of the warning.

Half of the participants in each group chose to issue an SVS on the Pontotoc storm (Fig. 5.5), and since no severe weather was reported for this storm, all updates were based only on maintenance reasons. The experimental group's statistically significant lower direction similarity scores were primarily because of E15's more careful adjustment of the warning polygon vertices and lack of editing within the text portion of the WarnGen AOI compared to other experimental participants. While the control group were more similar with respect to scanpath direction, they were statistically significantly less similar than the experimental group in the length and position MultiMatch dimensions (Fig. 5.6). The lower length similarity scores were due to participants C4 (Fig. 5.11c) and C15 (Fig. 5.11d) focusing their attention predominantly in the Reflectivity and WarnGen AOIs, respectively. C4 issued this update to trim the warning polygon, while C15 wanted to add text in the warning to communicate the expected hail threat. This result corresponds to C4 and C15 having the least fixations in the WarnGen and Reflectivity AOIs out of the control group, respectively. Finally, the statistically significant lower position scores in the control group were a result of participant C5's sporadically placed fixations that were likely caused due to his eye gaze darting between the keyboard and computer screen while editing warning text.

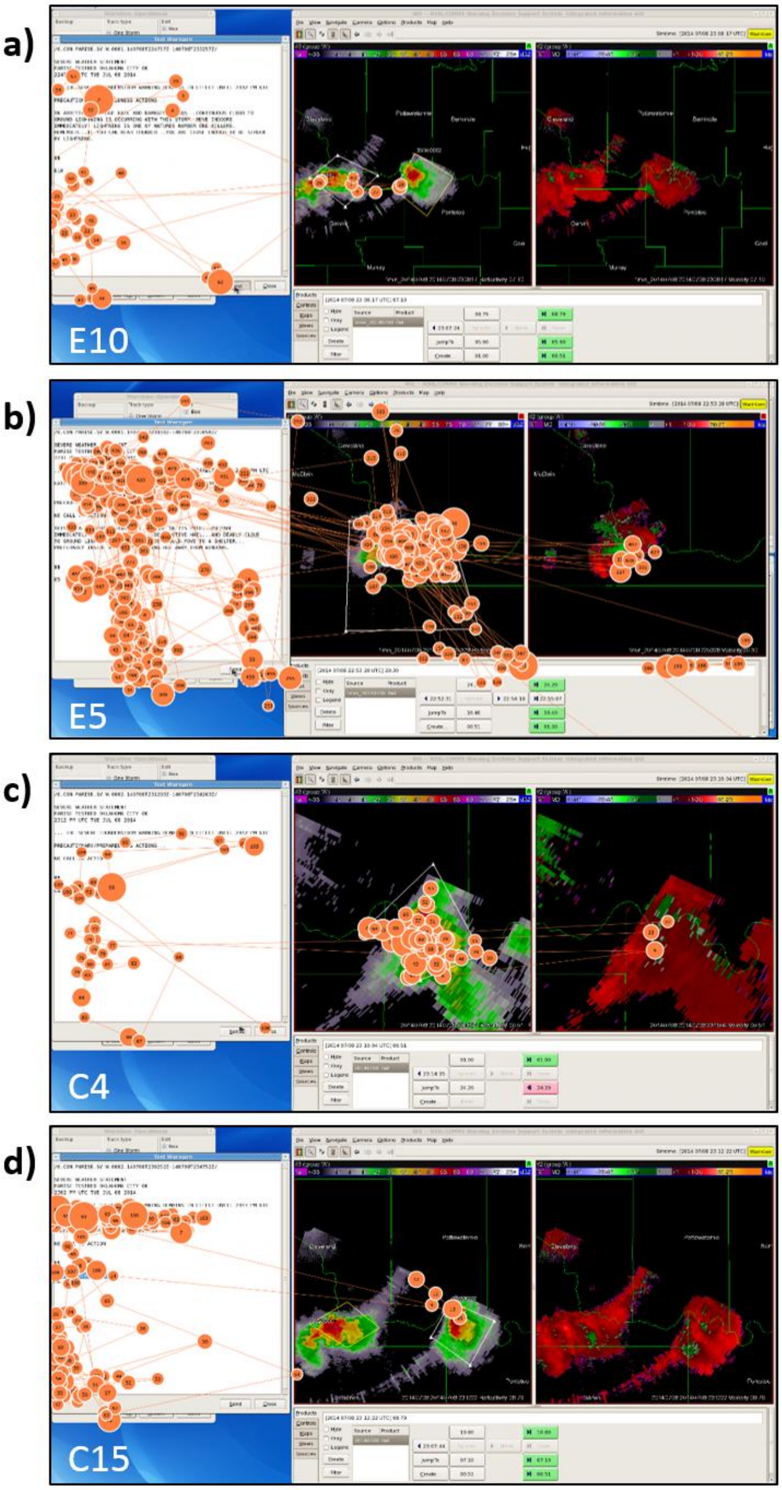


Figure 5.11. Gaze plots depicting the scanpaths of participants a) E10 and b) E5 during the first McClain storm warning update, and of participants c) C4 and d) C15 during the first Pontotoc storm warning update.

5.3.5 Differences in Duration

Unlike the vector, direction, length, and position MultiMatch measures, similarity in fixation duration is difficult to visualize in gaze plots and thus challenging to compare between forecasters. When focused on a piece of information, a person's fixation duration is indicative of their level of engagement and effort in extracting and processing it (Poole and Ball 2006; Bojko 2013). In each of the four defined stages, the difference in fixation duration similarity scores among control participants and among experimental participants was statistically significant at least once (Fig. 5.6e). However, in only one of these instances did the group with statistically significant *lower* duration similarity scores also have statistically significant *lower* similarity scores in other MultiMatch dimensions (Fig. 5.6). In the other instances, either no statistically significant difference was found for the vector, direction, length, or position dimensions, or the group that experienced statistically significantly *more* variation in duration was the one to experience statistically significantly *less* variation in other dimensions (Fig. 5.6). This result demonstrates that even when forecasters' placement of and transition between fixations is similar, how intently they focus on information can still vary.

5.4 Discussion and Summary

The hour-long scenario presented during this eye-tracking experiment provided an opportunity to collect forecasters' eye gaze data in a simplified warning scenario so that similarities and differences in their warning decision processes could be better identified. In this simplified warning scenario, the fixation measures of the control and experimental groups were generally similar throughout the four defined stages, with only a few statistically significant differences in mean fixation duration. The high degree of similarity in these measures is likely a result of two factors. First, all forecasters were asked to maintain focus on the screen throughout the event, and the comparable totals in fixation measures is therefore somewhat expected. However, forecasters still had the freedom to distribute their attention wherever they chose. We saw that the distribution of attention was generally similar among most participants regardless of whether 1-min or 5-min PAR updates were used. In retrospect, we believe that the chosen weather scenario strongly influenced this result. Prior to beginning the case, the majority of forecasters believed that the expected weather threat was primarily hail and secondarily wind. Given the minimal data available for interrogation, it is then unsurprising that forecasters focused predominantly in the Reflectivity AOI, switching often between the persistent McClain and Pontotoc storms, with more intermittent checking in the Velocity AOI. The threat expectation for these slow-moving multicell storms did not change throughout the case, and this interrogation pattern was thus maintained for much of the hour.

It remains to be seen whether differences in the fixation measures of forecasters using 1-min and 5-min PAR data would be greater if presented with a more complex weather event. The pilot study that motivated the use of eye tracking in this larger experiment observed a response in a forecaster's eye gaze data when the expected weather threat switched from severe hail to severe downburst winds (Wilson et al. 2016). Furthermore, the traditional experiment component of the 2015 PARISE showed that the impact of using rapidly-updating PAR data depends on the type of weather event presented (Wilson et al. 2017). For the case chosen in this study, while control participants expressed that faster PAR updates would have been useful to observe trends in more detail, their general attitude was that "…the storm changed slowly enough that not having the rapid update data wasn't a killer" (C2). It is possible then that a more dynamically evolving weather event may yield different results.

Forecasters' interrogation behaviors were further analyzed using the MultiMatch scanpath comparison algorithm (Jarodzka et al. 2010). Given that this algorithm considers eye movements on a much finer scale than the AOI-based fixation measures, this method better captured similarities and differences between how forecasters' fixations traversed the screen. Variability in scanpath behavior was found to be comparable within the two groups for all five MultiMatch dimensions only prior to the first McClain storm warning and during the issuance of the Pontotoc storm warning. For all other portions of the defined stages, either the control or experimental group was found to have statistically significantly more variation in at least one of the five MultiMatch dimensions. Examination of the video footage and forecasters' retrospective recall, as well as closer inspection of the similarity scores, revealed why this greater variability occurred. We did not find evidence that supported a direct link between scanpath similarity scores and participants' use of 1-min or 5-min PAR updates for this case. However, examples illustrated that the scanpath comparison results were useful for identifying participants who deviated away from the normal tendencies of a group. These deviations occurred because of how participants interacted with the user interface, tackled technological glitches in the WarnGen system, or approached tasks differently based on their understanding and expectations of the weather event.

The sensitivity of the MultiMatch scanpath comparison algorithm to differences in forecasters' behavior suggests that application of eye-tracking methods could be useful for exploring other avenues of operational meteorology research. In addition to testing forecasters' interactions with rapidly-updating radar data for other weather scenarios, eye tracking could be used to investigate how forecasters acquire and integrate other types of information into warning operations. With the fairly recent polarimetric upgrade to the WSR-88D network and the launch of the GOES-R series (Schmit et al. 2017), there are plenty of new data and experimental products to be tested through the lens of a weather forecaster. Eye movement data could also be used to validate models of forecasters' attention systems that are specifically designed to support their allocation of limited perceptual and cognitive resources when interrogating meteorological information (Schvartzman et al. 2017). Furthermore, given the successful applications of eye tracking in usability studies, these methods could be used to support the development and testing of user-friendly interfaces that display information in an efficient and effective manner to forecasters.

This study demonstrates how eye-tracking methods can be used to address operational meteorology research questions and will help inform future work that also intends to explore this research avenue. Based on this study alone we have learned that analyzing eye gaze data beyond the bulk measures of fixation count and duration is necessary for detecting differences in eye movement behavior. As we found in this study, analysis of participants' scanpaths are especially beneficial in scenarios that constrain the amount of content available to the participant and collect eye gaze data for a fixed duration (which can force similar fixation measure totals). The scanpath comparisons computed with the MultiMatch algorithm were well-suited for determining if, and in what ways, forecasters' sequences of fixations differed. Also noteworthy is that without examination of the qualitative data, making sense of the scanpath similarity scores would have been extremely difficult. We therefore emphasize the importance of collecting these data (i.e., video or retrospective recall) alongside the eye gaze data to aid contextual interpretation of forecasters' behaviors. Finally, through carefully designed experiments that obtain interpretable and meaningful data, we are hopeful that future eye tracking studies will expand our understanding of forecasters' cognition and act to support their important role within the weather enterprise.

Chapter 6

Considerations for Phased-Array Radar Data Use within the National Weather Service

Taken in full from: Wilson, K. A., P. L. Heinselman, and C. M. Kuster, 2017: Considerations for phased-array radar data use within the National Weather Service. *Wea. Forecasting,* in press.

Abstract

Thirty NWS forecasters worked with 1-min, 2-min, and 5-min PAR volumetric updates for a variety of weather events during the 2015 PARISE. Exposure to each of these temporal resolutions during simulated warning operations meant that these forecasters could provide valuable feedback on how rapidly-updating PAR data impacts their warning decision processes. To capture this feedback, forecasters participated in one of six focus groups. A series of open-ended questions guided focus group discussions, and forecasters were encouraged to share their experiences and opinions from the experiment. Transcriptions of focus group discussions were thematically analyzed and themes belonging to one of two groups were identified: 1) forecasters' use of rapidly-updating PAR data during the experiment, and 2) how forecasters envision rapidly-updating PAR data being integrated into warning operations. Findings from this thematic analysis are presented in this paper, and to illustrate these findings from forecasters' perspectives, dialogue that captures the essence of their discussions is shared. The identified themes provide motivation to integrate rapidly-updating radar data into

warning operations, and highlight important factors that need to be addressed for successful integration of these data.

6.1 Introduction

PARISE has completed four main studies to measure the impacts of rapidly-updating PAR volume scans on NWS forecasters' warning performance and related warning decision processes during a variety of weather events (Heinselman et al. 2012; Heinselman et al. 2015; Bowden et al. 2015; Bowden and Heinselman 2016; Wilson et al. 2017). In previous studies, forecasters were exposed to only 1-min or 5-min PAR updates. Although these studies demonstrated positive impacts of 1-min PAR update use on forecasters' situational awareness, applications of conceptual models, and accuracy and timeliness of warnings (e.g., Heinselman et al. 2015; Bowden et al. 2015), forecasters' experiences were constrained to a single temporal resolution of radar data.

The 2015 PARISE was unique in that all 30 participating NWS forecasters were exposed to three temporal resolutions of PAR volumetric updates. The opportunity to actively work with multiple radar update speeds meant that these forecasters were positioned to provide well-balanced feedback on what they considered to be the operational impacts of rapidly-updating PAR data. This feedback is important for informing future technology decisions and ensuring that their needs as users will be met should rapidly-updating radar data become a reality in future warning operations. Six focus groups were therefore conducted to enable forecasters to share their feedback and offer valuable insight from the 2015 PARISE.

6.1.1 Experiment Description

In the most recent PARISE, 30 NWS forecasters were each invited to participate in one week of the experiment, which took place in the NOAA Hazardous Weather Testbed over six weeks during August and September 2015. The experiment week that participants were assigned to only depended on their availability. The participants were recruited from 25 forecast offices located in the Great Plains and their forecasting experience ranged from 1 to 27 yr (mean = 12 yr, standard deviation = 7 yr). Throughout the week, forecasters worked a series of nine weather events, of which three were considered null, three presented severe hail and/or wind threats, and three presented tornado threats. The duration of each simulation ranged from 19–65 minutes. Forecasters were asked to independently interrogate reflectivity, velocity, and spectrum width products in simulated real-time and issue severe thunderstorm and tornado warning products as they considered them necessary. For each case, forecasters were provided with either 1-min, 2-min, or 5-min PAR volumetric updates depending on their random assignment to one of three groups. All groups rotated through each temporal resolution for the three null events, three severe hail and/or wind events, and three tornado events (see Wilson et al. 2017 for further details).

6.1.2 Focus Group Description

At the end of each of the six experiment weeks, a focus group was conducted that consisted of five participating forecasters, all of whom were from different forecast offices. Given that the focus group was the final activity of the week, both forecasters and researchers had already established rapport, thus encouraging honest and fruitful discussions. The focus groups were guided with a set of predetermined open-ended questions so that forecasters' responses were unconstrained (Lazar et al. 2010). These questions were specifically designed with a goal to elicit feedback on: forecasters' reactions and responses to the three temporal resolutions of PAR data, how these data affected their conceptual understanding of different weather events, and how they envision using these data in a real-time operational environment (see Appendix A for list of questions). Although the flow of discussion differed for each focus group, all participants were asked the same set of questions and discussions lasted between 1.5 to 2 hours. An advantage of collecting forecasters' feedback within a focus group setting was that interactions between participants helped create a synergistic effect, which in turn promoted the sharing of opinions and generation of ideas (Cameron 2010; Krueger and Casey 2015).

In this article, we present the findings from the analysis of forecasters' feedback. Transcriptions of the six focus group discussions were thematically analyzed according to their semantic content (Clarke et al. 2015). A list of codes was first developed to describe the content, and these codes were then reduced to a set of themes that belonged to one of two groups (Fig. 6.1). Given the qualitative nature of focus groups, findings related to the identified themes are expressed in impressionistic terms and are based solely on the viewpoints of forecasters participating in this study (Cameron 2010). To ensure anonymity in direct quotes, forecasters were assigned participant numbers P1–P30. This article describes each of the identified themes and shares the most inclusive and pertinent topics that forecasters discussed.

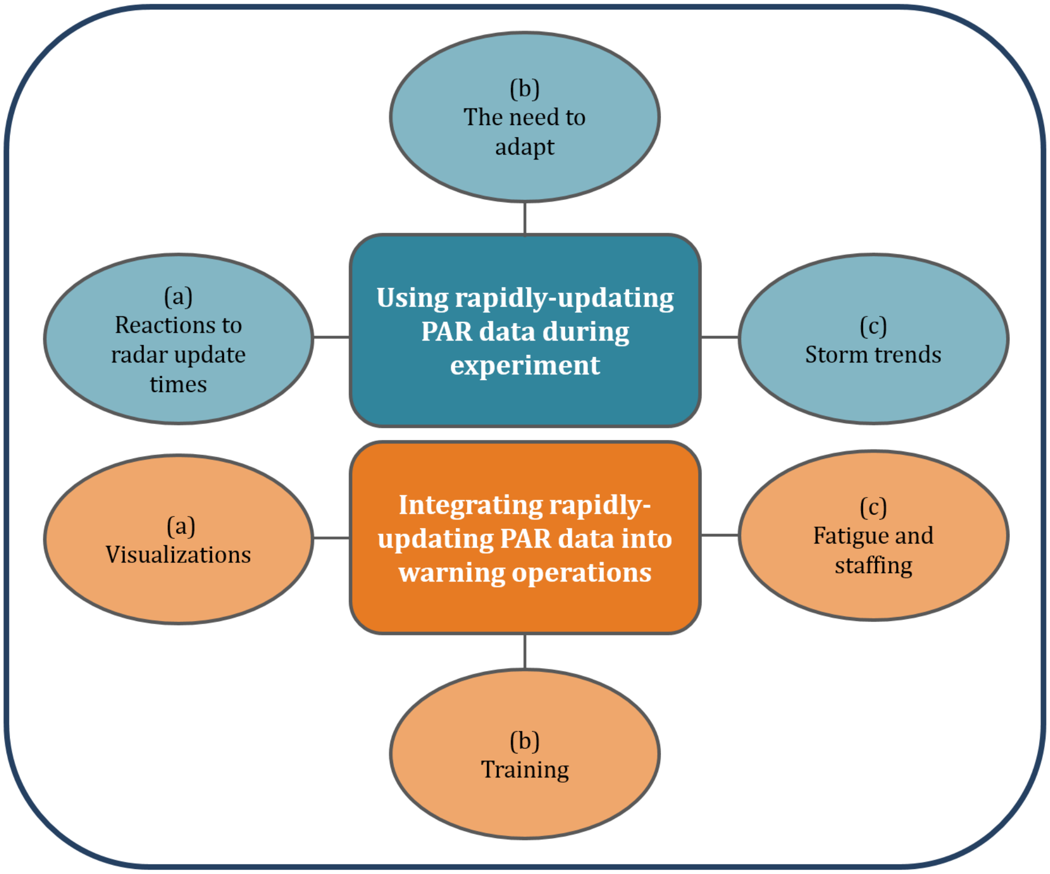

(b)
The need to adapt
(a)
Reactions to radar update times
Using rapidly-updating PAR data during experiment
(c)
Storm trends
(a)
Visualizations
Integrating rapidly-updating PAR data into warning operations
(c)
Fatigue and staffing
(b)
Training

Figure 6.1. Two groups of themes identified in transcriptions from forecasters' discussions during focus groups.

6.2 Using Rapidly-Updating PAR Data during the Experiment

6.2.1 Reactions to Radar Update Times

For all participating forecasters, their first opportunity to use rapidly-updating PAR data to make warning decisions was during this experiment. Describing their initial reactions to these data, forecasters focused on 1-min PAR updates and exhibited positive and upbeat attitudes because of their ability to now view how storms were evolving on shorter timescales. General statements were made, such as *"It was awesome. I know this is happening, but I can't see it with the 88D data. You miss everything in between"* (P21). Some forecasters also likened these data to textbook examples of storm processes, and pointed out that *"With the one-minute data it looks more like what you see when you are out in the field"* (P5).

Forecasters viewed these faster updates for three of nine cases that were worked in a randomized order and became used to the additional radar data very quickly. As P27 reported, their randomized case order meant that they worked three weather events with 1-min PAR updates first. P27 noted that that they "*…got used to the fast data fast,"* such that returning to 5-min PAR updates "*...Killed me…. It was like walking through wet cement and I wanted faster data."* Though the case order for other forecasters did not accentuate the difference between 1-min to 5-min PAR updates as much, they still became accustomed to the faster updates quickly, making statements that they were *"...waiting for data when I had slower data"*(P8), which *"...was like watching paint dry"* (P24). Thinking about their return to the forecast office, P27 said that they *"...can already tell that this is going to kill me during my first radar shift. I will just want the [faster] data!"*

Another point of discussion regarding forecasters' reactions to faster radar updates was how their sense of time became skewed. One participant pointed out that *"You see a new scan and think it has been five minutes,"* (P27) while another noted that *"With one-minute [updates], time seemed like it was going faster than it actually was"* (P9). Forecasters evidently use radar updates as an external cue for time progression during warning operations, and were either unaware of how strong this external influence is on their sense of time or were not actively prepared to shift their sense of time during this experiment.

6.2.2 The Need to Adapt

Despite forecasters being excited about the use of 1-min PAR updates to make warning decisions, approximately one third of participants reported feeling overwhelmed at first. This feeling resulted from trying to *"…keep up with everything coming in"* (P19) and *"…look at all tilts of everything"* (P15) at the same rate that the faster updates were being received. These participants reported that they soon realized interrogating faster updates in this manner *"…was not going to be possible"* (P15). P8 explained that *"It was nice to see all of the data, but to not become overwhelmed you had to quickly go through stuff and decide what you actually wanted to look at."* Forecasters therefore described needing to use a *"mental filter"* (P11) that was dependent on *"…the threat type and what your expectations are"* (P25) to better manage the increased amount of radar data. Applying a mental filter was most necessary during weather events that posed a tornado risk. Like many other forecasters, P2 explained that they *"Pushed hail aside and just watched 0.5 velocity like a hawk"* believing that it was *"…worth the trade off since you need to know about the tornado."* However, several participants cautiously added that this prioritization in attention should depend on the seriousness and location of threats. For example, P3 pointed out that *"If there is softball size hail over a town, you need to be looking aloft for the hail cores. Especially if the tornado is weak and in a rural area and the big hail or wind is in a town."* Therefore, focusing interrogation according to the primary threat may not always be an ideal solution for comfortably managing faster radar updates.

6.2.3 Storm Trends

When discussing the specifics of cases worked, forecasters focused heavily on their newfound ability to observe storm trends in much greater temporal detail when using faster radar updates. These forecasters explained that they *"Have more confidence when you can see evolutionary changes [because] you see what you are expecting to see, or maybe what you were not expecting to see"* (P20). Many of their shared examples from the experiment corroborated findings from earlier PARISE studies and drew on some of the previously reported sampling limitations of the WSR-88D (LaDue et al. 2010). For example, in pulse-type storm environments, forecasters appreciated being able to better observe the persistence of updrafts as well as track the development and location of high-reflectivity cores. Like others, P27 thought that "*…it was really cool to see the new updrafts form aloft. It was awesome to have fast data there. With five minute data a storm could pulse up and you won't even see it. So you could see your conceptual model evolve over time instead of making assumptions."* Similarly, P9 said that *"You can see so many more features. You can see the high reflectivity cores grow elevation scan to elevation scan. With the 88D it just shoots up, you know it increases, but you don't get to see it happen."* Additionally, being able to see hail cores *"descend minute by minute down to the surface"* aided forecasters in modifying the expected weather threat after a warning was issued, allowing them to *"put out an update and call for bigger hail"* (P23).

Forecasters also described the usefulness of faster radar updates for making tornado-related warning decisions during this experiment. In simulated warning operations, viewing radar indicated evidence of tornadogenesis in finer temporal detail has resulted in the issuance of earlier warnings by up to 7.5 minutes, especially during classic supercell events (Heinselman et al. 2015; Wilson et al. 2017). In the 2012 PARISE, forecasters achieving above-average tornado warning lead time applied conceptual models that depended on trends only observable in 1-min PAR updates (Heinselman et al. 2015). P13 emphasized the importance of these trends, reporting that *"I've never seen such a clear example of tornadogenesis in radar data before. You see the rear-flank downdraft kicking out, the midlevel meso dropping down. You saw what you would expect to see based on the textbook conceptual model. You could not see that with five minute data. I am confident that this allowed me to put a warning out sooner than with five minute data."* Despite these encouraging results, the most recent PARISE also found that extending tornado warning lead time through the use of faster radar updates was difficult to achieve for a weak and short-lived tornado that developed in a quasi-linear convective system (Wilson et al. 2017). Based on their use of rapidly-updating radar data for this single event, some participants explained that while 1-min PAR updates allowed them to observe brief circulations, it was unlikely that they would issue a tornado warning. Some forecasters reasoned that *"The fastest you can issue a warning is a minute or so, and by then the warning is out and not much is happening"* (P18). Nevertheless, forecasters did state that being able to observe these circulations was still beneficial for providing additional threat information in a severe thunderstorm warning.

Observing more-detailed storm trends was also helpful in preventing the issuance of warnings on storms that did not become severe. Forecasters reported being able to see that storms *"Never really got a great updraft for what I would think is needed to get a good downburst"* (P10) and that cores *"…Were not sustaining themselves for very long"*

(P7). While these observations did help reduce the number of false alarms in the 2015 PARISE (Wilson et al. 2017), a handful of forecasters noted that *"You have to be careful with how quickly you react to the one-minute data too"* (P11). Several forecasters using faster radar updates were disappointed in impulsive warning decisions made after viewing intensifications in storm trends that were only transient, and therefore recommended waiting to view consistency in trends before acting on them.

6.3 Integrating Rapidly-Updating PAR Data into Warning Operations

6.3.1 Visualizations

To create a mental image of storm structure and trends, forecasters currently analyze the vertical profile of storms in separate elevation scans and step back and forth in time to assess the temporal changes. This approach was found to be time consuming when using 1-min PAR updates during the experiment, leaving some forecasters feeling overwhelmed and many needing to limit their attention to portions of the storm that they believed posed the greatest threat. Forecasters identified that *"The answer to data overload might be integration in a 3D display, like GR or FSI… With the 5-minute data, I don't feel like cross sections or volume data is really that helpful, but with this data I could really see myself using those types of tools"* (P27). Furthermore, forecasters want trends to be monitored using an automated technique. P18 suggested that *"If AWIPS could somehow track a core and tell you how much the reflectivity is changing from scan to scan, you don't have to look and calculate for yourself. Something could tell you that the reflectivity has increased by 40 dBZ."* This idea would reduce the manual search efforts and corresponding demand on working memory for tracking trends, and could be extended to monitoring additional aspects of reflectivity cores, as well as the evolution of other precursor signatures.

6.3.2 Training

Forecasters drew on their experience of using faster radar updates to make recommendations for the type of training they would find most helpful, and unanimously agreed that hands-on experience is most valuable. Though not possible in PARISE, forecasters felt it would be advantageous to work weather events multiple times with different temporal resolutions of radar data. This activity would allow them to better assess how faster radar updates can benefit their warning decision process. As P25 pointed out, *"I don't know what I missed between scans."* Given that you *"Can see a lot of new processes"* (P2) that were previously unobservable, some forecasters suggested that providing a list for when faster radar updates are most beneficial to the warning decision process would also be helpful. Furthermore, forecasters noted that the greater temporal detail in storm processes will require them to revisit and possibly modify their conceptual models. One forecaster suggested that showing *"…Video of the storm alongside the radar so people can get used to seeing how the storm evolves and what that looks like on radar"* (P25) would aid this process, while another noted the importance of interrogating faster radar updates using *"Only base data without algorithms [to]…force you to go back to conceptual models"* (P8). Forecasters suggested completing hands-on training away from the forecast office in a set-up similar to the Warning Decisions Training Division's Radar Applications Course. This idea was preferable to within-office training because *"There are many more distractions"* (P29) within the forecast office and *"Sometimes it takes two months just to get everyone in the office through one case"* (P3).

Recognizing that resource limitations may make this idea difficult to execute, one feasible suggestion was that *"…You need to train the trainer. Bring one person from each office and then have them go back and teach the office"* (P4). The NWS Training Center recently adopted this strategy for the GOES-R preparation course, in which Science and Operations Officers and Development and Operations Hydrologists developed knowledge and experience that could then be shared with forecasters at the local level (personal communication, Brian Carcione).

In addition to receiving hands-on training, forecasters thought that step-by-step reviews of their own warning decision processes would be a useful training activity. As part of this experiment, forecasters were asked to watch a playback video of their onscreen activity and recall what they were seeing, thinking, and doing. P22 suggested *"What if at every office you sat people down and asked them what they were thinking minute by minute. Maybe we can improve what you are doing… That was helpful for me, since I have never been asked this before."* While most other forecasters agreed with this statement, a few felt that reviewing onscreen activity in this manner might make others feel as though their warning decisions are being judged. Importantly though, P25 emphasized that *"We need to be more thick skinned as a weather community with case reviews. What we have done this week is one step away from what an NFL team does each Monday when they dissect game film."* Forecasters therefore recognized that this review procedure would be a useful training approach for strengthening the performance of both the radar operator and forecast team as they learn to integrate faster radar updates into warning operations.

6.3.3 Fatigue and Staffing

Ensuring that humans are operating within their optimum working conditions is important for both their well-being and their performance. While cognitive workload associated with the use of rapidly-updating PAR data has been assessed within the PARISE setting (Wilson et al. 2017), it has not been measured in live operations where forecasters are part of a team and are exposed to many other data sources. Some forecasters expressed their concern of the "fatigue factor," where *"It would be a bigger factor with rapid-update data since you are interrogating more data… We are already concerned about that. We talk about it every spring. How long are we going to let someone look at radar data? With new types of radar data, that conversation is important again"* (P21). Reflecting on this matter, forecasters stressed that to work efficiently with faster radar updates and to ensure smooth function of warnings operations, they would need to redistribute responsibilities within their teams. Forecasters expect that "*There will be an increasing need to sectorize"* (P17), meaning that *"There will need to be more radar operators"* (P9). Additionally, forecasters recommended sharing the task of updating warnings so that the radar operator could focus on issuing warnings only.

6.4 Discussion and Conclusions

Communicating findings from the focus group discussions gives forecasters a voice in the research process and allows for an evaluation of rapidly-updating PAR data from their specialized perspectives. The six focus group discussions brought to attention the ways in which forecasters felt these data benefited their warning decision processes and highlighted some important considerations that need be addressed should these data be implemented operationally.

The consensus among forecasters was that 1-min PAR updates are preferable to 2-min and 5-min PAR updates. This preference was further evident in their choice to emphasize and share experiences that were predominantly related to their use of 1-min PAR updates during the experiment, with only little to no attention given to their use of 2-min PAR updates. Forecasters' lack of comments regarding 2-min PAR updates was surprising given other forecasters' suggestions in earlier studies that it would be helpful to show 2-min PAR updates in addition to 1-min PAR updates (Bowden and Heinselman 2016). However, capturing the feelings of others, P25 summarized that *"At the end of the day, radar data is at the heart and soul of warning operations. If it stops, you are severely handicapped. So, it is critically valuable, especially if it is one minute, because it is giving you a constant idea of what the storms are doing and where the storm is and where it is moving and where it has been. It has to be integrated in some way, shape, or form."*

Despite strong consensus that forecasters preferred the use of 1-min PAR updates, some disagreement in how to manage these data emerged in the focus group discussions. First, while numerous forecasters thought that the development of new algorithms could provide a solution to the increased levels of workload associated with tracking 1-min trends in radar signature, others expressed concern that forecasters might become dependent on these algorithms and lose their sense of conceptual understanding. Second, many forecasters found that prioritizing attention to the primary severe weather threat helped counteract high levels of workload. However, several forecasters thought that this approach was not suitable for dealing with scenarios that presented multiple weather threats. Future research efforts should examine the feasibility of these suggested solutions in an experimental setting where the impacts of algorithm use and prioritization of attention on forecasters' warning decision processes can be assessed independently.

In addition to forecasters' suggestions of employing new strategies for viewing 1-min radar updates, being able to successfully alleviate the inevitable increase in radar operator demands will depend on the ability of forecast office staff to redistribute responsibilities. During the 2015 PARISE, forecasters reported experiencing levels of high and excessive workload more frequently when using 1-min PAR updates during events that presented a tornadic threat (Wilson et al. 2017). Oftentimes, this spike in cognitive workload occurred during times in which forecasters were issuing or updating a warning, which led to forecasters' recommendation that sharing product issuance tasks among multiple radar operators would be one helpful approach to decreasing cognitive load. Furthermore, during weather events that are more demanding on forecasters' attention, the presence of multiple radar operators would be beneficial for sectorizing warning area and reducing individual forecaster's overall task load.

Forecasters' positive attitudes and outlooks of using rapidly-updating PAR data within the forecast office are encouraging. Successful implementation of rapidly-updating radar data will first require the delivery of hands-on training. Because logistical limitations will likely prevent all forecasters from completing a course at a training center location, an approach similar to the GOES-R preparation course is recommended. In this instance, specific individuals from forecast offices receive specialized training and transfer their learned knowledge and skills to other forecasters upon their return. Additionally, given that many forecasters commented on the usefulness of completing retrospective recalls during the 2015 PARISE, we believe that adopting this practice as a form of training will enhance forecasters’ capacities to understand and improve upon their own warning decision making behavior. Although some forecasters expressed frustration at finding time to complete training during work hours, in-house training must become a priority to ensure a smooth transition to using rapidly-updating radar data in warning operations.

Although forecasters have not yet used 1-min PAR volumetric updates during real warning operations, their use of the recently implemented MESO-SAILS scanning strategy could provide some interesting insight for the potential integration of PAR data in the future (Chrisman 2014). MESO-SAILS allows forecasters to receive up to three additional interspersed 0.5° elevation scans during a volumetric update. While this scanning strategy does not mimic the rapid updates that PAR obtains for the entire volume scan, a review of the initial impact of these more frequent low-level observations on forecasters’ warning performance should be completed. This review would be a first step to investigating some of the focus group findings in real-time operations. Forecasters indicated that responding too quickly to transient trends in radar signatures could negatively affect their warning decisions. Assessing forecasters’ use of MESO-SAILS within operations with respect to their reactions to trends viewed in the 0.5° elevation scan would thus be worthwhile. Additionally, given that forecasters in the focus group described experiencing a skewed sense of time while interrogating 1-min PAR updates, it would be interesting to explore whether forecasters using MESO-SAILS within the naturalistic environment also need to modify their sense of time when consistently tracking the 0.5° elevation updates. Finally, important lessons could be gained from investigating the overall implementation of MESO-SAILS into the forecast office, the preparations that forecasters found helpful prior to their use of these additional data, and how they adapted their interrogation styles to effectively incorporate these data into their warning decisions processes.

Chapter 7

Conclusions, Implications, and Future Work

Weather radar data, as one NWS forecaster described during the 2015 PARISE, "is at the heart and soul of warning operations." Radar allows forecasters to observe the dynamic nature of potentially hazardous weather and to alert those at risk through the issuance of weather warnings. Since PAR technology, which can provide faster volumetric updates, is being considered as a potential replacement for the current WSR-88D system (Zrnić et al. 2007; Stailey and Hondl 2016), examining the potential impacts of rapidly-updating radar data on forecasters' warning performance and related warning decision processes is essential.

The 2010, 2012, and 2013 PARISE studies reported that forecasters' use of rapidly-updating radar data resulted in predominantly positive impacts on their application of conceptual models, their overall situational awareness of weather events, and on their warning lead time and accuracy statistics (e.g., Heinselman et al. 2012, 2015; Bowden et al. 2015). However, the chosen methodologies limited the generalizability of these findings and important research questions remained unexplored. The 2015 PARISE traditional experiment addressed the sample size limitations of earlier PARISE studies. With an increased number of participating forecasters and an increased number of cases worked, forecasters' overall performance in the 2015 PARISE supports previous findings that median warning lead time increases with use of increasing radar update speed (Wilson et al. 2017). Additionally, forecasters' use of these data resulted in fewer false alarms and an enhanced ability to discriminate correctly between weather threats compared to forecasters who were provided traditional 5-min radar updates.

Despite these encouraging results, the use of rapidly-updating radar data did not improve warning performance in all cases. The most surprising result was that almost all forecasters failed to achieve positive warning lead time for a short-lived tornado that occurred within a bowing line segment. Although these events are notoriously difficult to warn for, our expectation was that forecasters' use of 1-min or 2-min PAR updates would improve their ability to provide warning lead time for this tornado. Given that only a single case of this event type was analyzed during the 2015 PARISE, future research should investigate how benefits of rapidly-updating radar data can be realized for improving warnings for nonsupercell tornadoes. Discussions with forecasters suggest that this research should analyze both their radar interrogation strategies and the associated warning philosophies that are ingrained into forecast office practices for these event types.

In addition to analyzing forecasters' warning performance during the traditional experiment, forecasters' cognitive workload was assessed during the 2015 PARISE to investigate whether use of rapidly-updating radar data increases their susceptibility to experiencing excessive cognitive workload (i.e., "overload"). The results showed that, in general, forecasters' subjective ISA ratings were skewed towards higher levels of cognitive workload with use of faster radar updates, but that experiences of cognitive overload were rare (Wilson et al. 2017). Forecasters provided reasoning along with the ISA ratings that brought to light why, aside from their use of rapidly-updating radar data, their cognitive workload increased to high and excessive levels. Forecasters' reasoning revealed that storm characteristics, warning tasks, beginning a case in experimental conditions, technical frustrations, and personal needs led to increased levels of cognitive workload. Understanding the interplay of these influencing factors during warning operations will help identify strategies that support a radar operator's use of additional radar data, thus stabilizing their cognitive workload and ensuring their optimal performance.

Aside from PARISE, the 2016 Probabilistic Hazard Information experiment is the only other known research project that has specifically examined forecaster cognitive workload (Ling et al. 2017). Given that introducing other types of additional data and products to forecasters will likely modulate their cognitive workload, it would be valuable if researchers conducting forecaster-oriented experiments also considered documenting this aspect of forecasters' experiences. For research meteorologists unfamiliar with human factors methods, the cognitive workload assessments presented in the 2015 PARISE and the 2016 Probabilistic Hazard Information experiment provide helpful guidance on how to address this type of research question.

The second component of the 2015 PARISE successfully implemented eye-tracking research methods in a large-scale simulated real-time experiment for the first time. Given the limitations of the retrospective recall method (i.e., inaccuracies, incompleteness, and biases), an eye-tracking system was used to collect objective data on the distribution of forecasters' visual attention and related warning decision processes. These data were used to assess differences in how forecasters interacted with the radar display and warning interface. The MultiMatch algorithm proved to be useful for identifying and quantifying differences in forecasters' visual scanpaths, and the video and retrospective recall data were important for determining why these differences occurred. Forecasters deviating from their group's typical scanpath patterns were found to approach tasks differently, encounter more technological problems, or have greater variation in how they interacted with the user interface. However, differences in forecasters' scanpaths and related warning decision processes were not found to be directly associated with the temporal resolution of radar data that they used.

In hindsight, it is possible that the chosen weather scenario resulted in an overall lack of dissimilarity between the visual scanpaths of forecasters using 1-min radar updates and forecasters using 5-min radar updates. Most forecasters began working the case with an expectation that the weather threat was primarily hail and secondarily wind, and forecasters therefore applied a typical interrogation strategy that focused mostly in the Reflectivity AOI. Given that this eye-tracking experiment was the first of its kind, additional cases need to be tested in future work to assess whether the temporal resolution of radar data impacts forecasters' visual scanpaths during more challenging weather events. Furthermore, simplifying the design of the eye-tracking experiment presented in this dissertation would allow for analysis of basic forecaster eye movement behavior. Simplification could be achieved, for example, by reducing the duration of the experiment substantially, removing forecasters' abilities to zoom and pan data within AOIs, and restricting when radar scans of different times and elevations can be viewed. Results from a simplified experiment of this nature could aid in the interpretation of forecasters' eye movements in more complex scenarios.

Importantly, the eye-tracking experiment provides an example of how to integrate eye-tracking technology into a testbed experiment design in the context of an operational meteorology study. The eye-tracking experiment also demonstrated what processes are required to collect eye movement data and what types of analyses are useful for meaningfully interpreting the data. Hopefully, the successful implementation of eye-tracking research methods during the 2015 PARISE will encourage other scientists studying the human aspect of weather forecasting to explore eye-tracking applications within their areas of specialty and contribute towards an improved understanding of forecasters' cognition. For example, eye-tracking research methods could be applied to learn about forecasters' use of other types of meteorological data, such as satellite data, numerical weather prediction, or probabilistic hazard information. Using a similar approach to the 2015 PARISE eye-tracking experiment, tracing how forecasters acquire and then use this information would result in an improved understanding for how to most effectively integrate these data into the warning decision process. These eye-tracking data can also be used to validate and score models of forecasters' attention systems (e.g., Schvartzman et al. 2017), such that the most salient-rich features identified in presentations of meteorological data can be compared to forecasters' actual distributions of attention. Additionally, forecasters' eye movements can be analyzed during goal-oriented tasks to test the usability of newly developed display interfaces, which can lead to recommendations that improve human-computer interactions for operational meteorologists (Jacob and Karn 2003).

In addition to eye-tracking, focus groups were also used for the first time in PARISE during the 2015 study. The focus groups gave forecasters a voice in the research process so that they could share how rapidly-updating radar data impacted their warning decision processes during the experiment and how they envision these data being integrated into future warning operations. Forecasters were in agreement that of the 1-min, 2-min, and 5-min radar updates used during the 2015 PARISE, 1-min PAR updates were always preferable. Recommendations were provided for ways to manage the increase in available data assuming 1-min radar updates become operational. Some recommendations include redistributing responsibilities within the forecast office, prioritizing attention, and using algorithms. While forecasters were in agreement with the first recommendation, it is important to acknowledge that successfully redistributing responsibilities will depend on the availability, willingness, and flexibility of staff within a forecast office — a luxury that can be difficult to come by in offices that are operating with limited resources and personnel. Although the latter two recommendations for reducing workload can be achieved without redistributing responsibilities, forecasters did not express complete agreement with these ideas. Some forecasters were concerned that prioritizing attention would result in forecasters missing the potential of a secondary weather threat, while others were worried that the use of algorithms would reduce forecasters' conceptual understanding of weather events. This disagreement highlights an area for future research that would assess the individual impacts of prioritizing attention and algorithm use on warning performance when forecasters are provided 1-min radar updates. Findings from this research would help determine whether these recommendations provide feasible solutions for managing additional radar data, or whether making adjustments to the work flow of an integrated warning team will be most effective for reducing the radar operator's workload.

Forecasters' feedback during the focus groups provided an operational perspective that was difficult to obtain within the traditional and eye-tracking experiments. Since the specific research questions investigated in these experiments could only be answered in simplified, controlled, and simulated warning operations, removing the normal functions, interactions, and nuisances of a forecast office was necessary. Although findings from the focus group discussions gave some insight into what could be expected upon initial implementation of rapidly-updating radar data into the forecast office, extending the work of PARISE to the naturalistic setting will be an important step for learning about the impacts of faster radar updates on everyday warning operations. The success of NWS WFOs' participation in past naturalistic studies (e.g., Hoium et al. 1997; Morss and Ralph 2007; Henderson et al. 2017) suggest that conducting research of this nature to further the efforts of PARISE would be a feasible and worthwhile endeavor. However, to successfully execute real-time use of new meteorological data in an NWS WFO, specific instrumentation, technical infrastructure, and data display capability is required.

While meeting real-time experiment requirements can prove challenging, ongoing projects within the meteorological community are well-positioned to begin exploring this research avenue. For example, the Engineering Research Center for Collaborative Adaptive Sensing of the Atmosphere's Urban Testbed Project is already working to provide and demonstrate the usefulness of higher-temporal resolution radar products to local stakeholders, including NWS forecasters (e.g., Chen and Chandrasekar 2015). Additionally, the Verification of the Origins of Rotation in Tornadoes Experiment-Southeast research program has set out to improve analysis and forecast systems, better understand how forecasters warn for southeastern tornadoes, and study how end users respond to forecast information (Rasmussen 2015). In addition to the use of mobile radars and other instruments during this research program, the existing collaborative effort between researchers, social scientists, and meteorologists makes this research program ideal for conducting naturalistic studies in local forecast offices.

In addition to analyzing rapidly-updating radar data within the operational environment, there are several other future research opportunities that have not yet been discussed. First, the anticipated installation of the Advanced Technology Demonstrator at the National Weather Radar Testbed in 2018 will be an important step towards developing a modern weather radar system that, like the WSR-88D, will have dual-polarization capability. The Advanced Technology Demonstrator will allow engineers to investigate the errors of PAR polarimetric variable estimates that are caused by differences in copolar antenna patterns between the horizontal and vertical polarizations (Ivìc 2017). Through this research and development, the analysis of rapidly-updating polarimetric radar signatures will be possible, and the knowledge gained from these analyses will help to inform scientific conceptual models of storms processes (Kuster et al. 2017a).

Although WSR-88D polarimetric data have shown to provide additional information about storm processes (Kumjian 2013), discussions with forecasters suggest that their current use of these data during the warning decision process is mostly confined to instances in which a confirmation of hazardous weather is sought (e.g., a tornado debris signature; Kumjian and Ryzhkov 2008; Bodine et al. 2014). These discussions therefore bring to question whether forecasters are utilizing the full benefits of polarimetric radar data during warning operation. A combination of observations through naturalistic study, interviews, and surveys with NWS forecasters would be useful for learning about how their understanding of these data, chosen strategies for analyzing information during the warning decision process, and training experience have influenced their current level of engagement with polarimetric radar data. Creating an awareness of the challenges that forecasters currently face when attempting to integrate polarimetric radar data into their warning decision processes will give guidance on how to ensure that these data can be of use to forecasters both prior to and after the issuance of a warning.

A second area for future PAR research relates to numerical weather prediction. Data assimilation experiments have shown that the use of higher-temporal resolution PAR data in convective-scale models significantly improves short-term forecasts, even more so for adaptively-scanned PAR data (Yussouf and Stensrud 2010; Supinie et al. 2017). While the recently funded Spectrum Efficient National Surveillance Radar (SENSR) program will support further data assimilation experiments with PAR data, these early results suggest that an operational PAR system that provides frequent updates has potential to improve storm-scale modeling. Furthermore, the Advanced Technology Demonstrator will allow for investigation into whether assimilating rapidly-updating polarimetric radar data further improves short-term forecasts. Given that the assimilation of WSR-88D polarimetric data has already shown to improve analyses and forecasts of convective storms (including their associated updrafts, reflectivity structures, and forecast updraft helicity tracks; Carlin et al. 2017), investigating how the temporal resolution of polarimetric data affects the outcome of data assimilation experiments will be an important next step.

These research efforts in numerical weather prediction are working to support the success of Warn-on-Forecast, a program which is exploring the possibility of 0–3 hour forecasts for high-impact weather with guidance from an on-demand, storm-resolving model forecast system (Stensrud et al. 2009). A major goal of Warn-on-Forecast is to use this probabilistic guidance to increase warning lead time out to one hour. This extended warning lead time is expected to be especially beneficial to end-users with specialist needs, such as schools and hospitals, who must actively make weather-related decisions ahead of when warnings are typically issued. To begin exploring how probabilistic guidance influences weather decisions, forecasters' use of the NSSL Experimental Warn-on-Forecast System for ensembles output was recently evaluated during the 2017 Experimental Forecast Program in the NOAA Hazardous Weather Testbed. However, further research is required to develop a better understanding of how forecasters can best utilize and communicate uncertainty information to stakeholders and the general public. Assessing the potential value of this information to non-NWS end users will help ensure that its benefits are realized not just by NWS forecasters, but by all members operating within the weather enterprise (e.g., Kuster et al. 2017b). Findings from this research will support the meteorology community's efforts to reinvent the watch and warning system and achieve the FACETs vision, in which a continuum of probabilistic information will instead drive forecasters' and end users' weather-related decisions.

References

Aaltonen, A. A., Hyrskykari, and K. Räihä, 1998: 101 Spots, or how do users read menus? *Proc. of CHI 98 Human Factors in Computing Systems*. 1998, Los Angeles, CA, ACM Press, 132–139.

Albert, W., 2002: Do web users actually look at ads? A case study of banner ads and eye tracking technology. *Proc. of the 11th Annual Conference of the Usability Professionals' Association*. 2002, Orlando, FL.

Al-Moteri, M. O, M. Symmons, V. Plummer, and S. Cooper, 2017: Eye tracking to investigate cue processing in medical decision making: A scoping review. *Computers in Human Behavior*, **66**, 52–66.

Anderson, N. C., F. Anderson, A. Kingstone, and W. F. Bischof, 2015: A comparison of scanpath comparison methods. *Behavior Research Methods*, **47**, 1377–1392.

Andra, D. L., E. M. Quoetone, and W. F. Bunting, 2002: Warning decision making: The relative roles of conceptual models, technology, strategy, and forecaster expertise on 3 May 1999. *Wea. Forecasting*, **17**, 559–566.

Argyle, E., J. Gourley, Z. Flamig, T. Hansen, and K. Manross, 2017: Towards a user centered design of a weather forecasting decision support tool. *Bull. Amer. Meteor. Soc.*, **98**, 373–382.

Atlas, D., 1976: Severe local storms. *Bull. Amer. Meteor. Soc.*, **57**, 398–435.

Atlas, D., and C. W. Ulbrich, 1990: Early foundations of the measurement of rainfall by radar. *Radar in Meteorology: Battan Memorial and 40th Anniversary Radar Meteorology Conference*. D. Atlas, Ed., Amer. Meteor. Soc., 86–97.

Austin, P. M., and S. G. Geotis, 1990: Weather radar at MIT. *Radar in Meteorology: Battan Memorial and 40th Anniversary Radar Meteorology Conference*. D. Atlas, Ed., Amer. Meteor. Soc., 22–31.

Beatty, J., 1982: Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, **91**, 276–292.

Benel, D. C. R., D. Ottens, and R. Horst, 1991: Use of an eye tracking system in the usability laboratory. *Proc. of the Human Factors Society 35th Annual Meeting*. Santa Monica, CA, Human Factors and Ergonomics Society, 461–465.

Bent, A. E., 1943: Radar echoes from atmospheric phenomena. Division 14 Report 42-2, Radiation Laboratory, CP-621.1-M1.

Berka, C., and Coauthors, 2004: Real-time analysis of EEG indices of alertness, cognition, and memory with a wireless EEG headset. *International Journal of Human-Computer Interaction*, **17**, 151–170.

Berka, C., and Coauthors, 2007: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, **78**, B231–B244.

Bertram, R., J. Kaakinen, F. Bensch, L. Helle, E. Lantto, P. Niemi, and N. Lundbom, 2016: Eye movements of radiologists reflect expertise in CT study interpretation: A potential tool to measure resident development. *Radiology*, **281**, 805–815.

Best, W. H., 1973: Radars over the hump: Recollections of the first weather radar network. *Bull. Amer. Meteor. Soc.*, **54**, 205–208.

Bodine, D. J., R. D. Palmer, and G. Zhang, 2014: Dual-wavelength polarimetric radar analyses of tornadic debris signatures. *J. Appl. Meteor. Climatol.*, **53**, 242–261.

Bojko, A., 2013: *Eye Tracking the User Experience: A Practical Guide to Research*, Rosenfield Media, 304 pp.

Bowden, K. A., P. L. Heinselman, D. M. Kingfield, and R. Thomas, 2015: Impacts of phased array radar data on forecaster performance during severe hail and wind events. *Wea. Forecasting*, **30**, 389–404.

Bowden, K. A, and P. L. Heinselman, 2016: A qualitative analysis of NWS forecasters' use of phased-array radar data during severe hail and wind events. *Wea. Forecasting*, **31**, 43–55.

Braham, R. R., 1948: *Thunderstorm structure and circulation. M.S. Dissertation, Dept. of Meteorology*, University of Chicago, 40 pp.

Brotzge, J., and S. Erickson, 2009: NWS tornado warnings with zero or negative lead times. *Wea. Forecasting*, **24**, 140–154.

Brotzge, J., and W. Donner, 2013: The tornado warning process: A review of current research, challenges, and opportunities. *Bull. Amer. Meteor. Soc.*, **94**, 1715–1733.

Brown, R. A., L. R. Lemon, and D. W. Burgess, 1978: Tornado detection by pulsed Doppler radar. *Mon. Wea. Rev.*, **106**, 29–38.

Brown, R. A., and J. M. Lewis, 2005: Path to NEXRAD: Doppler radar development at the National Severe Storms Laboratory. *Bull. Amer. Meteor. Soc.*, **86**, 1459–1470.

Brown, R. A., R. M. Steadham, B. A. Flickinger, R. R. Lee, D. Sirmans, and V. T. Wood, 2005: New WSR-88D volume coverage pattern 12: Results of field tests. *Wea. Forecasting*, **20**, 385–393.

Brown, R. A., and V. T. Wood, 2012: Simulated vortex detection using a four-face phased-array Doppler radar. *Wea. Forecasting*, **27**, 1598–1603.

Burgess., D. W., K. E. Wilk, J. D. Bonewitz, K. M. Glover, D. W Holmes, and J. Hinkelman, 1979: Doppler radar: The Joint Doppler Operational Project. *Weatherwise*, **32**, 72–75.

Buswell, G. T., 1935: *How People Look at Pictures: A Study of the Psychology of Perception in Art*, University of Chicago Press, 198 pp.

Byers, H. R., 1970: Recollection of the war years. *Bull. Amer. Meteor. Soc.*, **51**, 214–217.

Byers, H. R, and R. R. Braham, 1948: Thunderstorm structure and circulation. *Journal of Meteorology*, **5**, 71-86.

Byers, H. R., and R. R. Braham, 1949: The Thunderstorm. U.S. Govt. Printing Office, Washington DC, 287 pp.

Byrne, M. D., J. R. Anderson, S. Douglas, and M. Matessa, 1999: Eye tracking the visual search of click-down menus. *Proc. of CHI 99. 1999*. Pittsburgh, PA, New York: ACM Press, 402–409.

Cain, B., 2007: A review of mental workload literature. Report RTO-TR-HFM-121-Part II, 34pp. [Available online at http://www.dtic.mil/dtic/tr/fulltext/u2/a474193.pdf]

Calhoun, K. M., T. M. Smith, D. M. Kingfield, J. Gao, and D. J. Stensrud, 2014: Forecasters use and evaluation of real-time 3DVAR analyses during severe thunderstorm and tornado warning operations in the Hazardous Weather Testbed. *Wea. Forecasting*, **29**, 601–613.

Call, D. A., 2009: An assessment of National Weather Service warning procedures for ice storms. *Wea. Forecasting*, **24**, 104–120.

Cameron, J., 2010: Focusing on the focus group. *Qualitative Research Methods in Human Geography*, I. Hay, Ed., Oxford University Press, 152–172.

Card, S. K., 1984: Visual search of computer command menus. *Attention and Performance X, Control of Language Processes*, H. Bouma and D. G. Bouwhuis, Eds., Psychology Press, 97–108.

Carlin, J. T., J. Gao, J. C. Snyder, and A. V. Ryzhkov, 2017: Assimilation of Z_{DR} columns for improving the spin-up and forecast of convective storms in storm-scale models: Proof-of-concept experiemnts. *Mon. Wea. Rev.*, in press.

Chen, Haonan, and V. Chandrasekar, 2015: The quantitative precipitation estimation system for Dallas–Fort Worth (DFW) urban remote sensing network. *Journal of Hydrology*, **531**, 259- 271.

Chrisman, J. N., 2009: Automated Volume Scan Evaluation and Termination (AVSET): A simple technique to achieve faster volume scan updates. Preprints, *34th Conf. on Radar Meteorology,* Williamsbug, VA, Amer. Meteor. Soc., P4.4. [Available online at https://ams.confex.com/ams/pdfpapers/155324.pdf.]

Chrisman, J. N., 2014: The continuing evolution of dynamic scanning. NEXRAD Now, No. 23, NOAA/NWS/Radar Operations Center, Norman, OK, 8–13.

Cinaz, B., B. Arnrich, R. Marca, and G. Tröster, 2013: Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and Ubiquitous Computing,* **17**, 229–239.

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

Clarke, V., V. Braun, and N. Hayfield, 2015: Thematic analysis. *Qualitative Psychology: A Practical Guide to Research Methods,* J. A. Smith, Ed., SAGE Publications Inc., 222–248.

Clement, J., T. Kirstensen, and K. Grønhaug, 2013: Understanding consumers' in-store visual perception: The influence of package design features on visual attention. *Journal of Retailing and Consumer Services*, **20**, 234–239.

Cooper, G. E. and R. P. Harper, 1969: The use of pilot rating in the evaluation of aircraft handling qualities. NASA Technical Note D-5153, 60 pp. [Available online at https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19690013177.pdf]

Cristino, F., S. Mathot, J. Theeuwes, and I. D. Hilchrist, 2010: ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, **3**, 692–700.

Crum, T. D., and R. L. Alberty, 1993: The WSR-88D and the WSR-88D operational support facility. *Bull. Amer. Meteor. Soc.*, **74**, 1669–1687.

Crum, T., S. D. Smith, J. N. Chrisman, R. E. Saffle, R. W. Hall, and R. J. Vogt, 2013: WSR-88D Radar Projects—Update 2013. *Proc. 29th Conf. on Environmental Information Processing Technologies*. Austin, TX, Amer. Meteor. Soc., 6B.1. [Available online at https://ams.confex.com/ams/93Annual/ webprogram/Paper221461.html.]

Da Silva, F. P., 2014: Mental workload, task demand, and driving performance: What relation? *Social and Behavioral Sciences*, **162**, 310–319.

Daipha, P., 2015: *Masters of Uncertainty: Weather Forecasters and the Quest for Ground Truth.* University of Chicago Press, 280 pp.

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

De Waard, D., 1996: The measurement of driver's mental workload. PhD dissertation, Traffic Research Center, University of Groningen, 135 pp. [Available online at http://apps.usd.edu/coglab/schieber/pdf/deWaard-Thesis.pdf]

Deubel, H., 2008: The time course of presaccadic attention shifts. *Psychological Research*, **72**, 630–640.

Dewhurst, R., M. Nsytröm, J. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist, 2012: It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, **44**, 1079–1100.

Dijkstra, E. W., 1959: A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.

Djamasbi, S., M. Siegel, and T. Tullis, 2010: Generation Y, web design, and eye tracking. *International Journal of Human-Computer Studies*, **68**, 307–323.

Dranidis, V. D., 2003: Backboards of the fleet: Shipboard phased-array radar: A survey of requirements, technologies, and operational systems. *Journal of Electronic Defense*, 26, **55**–62.

Drost, R., J. Trobec, C. Steffke, and J. Libarkin, 2015: Eye tracking: Evaluating the impact of gesturing during televised weather forecasts. *Bull. Amer. Meteor. Soc.*, **96**, 387–392.

Duchowski, A. T., 2002: A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, **34**, 455–470.

Duchowski, A., 2007: *Eye Tracking Methodology: Theory and practice*. London: Springer-Verlag, 327pp.

Ellis, S., R. Candrea, J. Misner, C. S. Craig, C. P. Lankford, and T. E. Hutshinson, 1998: Windows to the soul? What eye movements tell us about software usability. *Proc. of the Usability Professionals' Association Conference*. 1998, Washington, DC, Usability Professionals' Association, 151–178.

Emersic, C., P. L. Heinselman, D. R. MacGorman, and E. C. Bruning, 2011: Lightning activity in a hail-producing storm observed with phased-array radar. *Mon. Wea. Rev.*, **139**, 1809–1825.

Endsley, M. R., 1995: Toward a theory of situation awareness in dynamic systems. *Human Factors*, **37**, 32–64.

Erickson, S., and H. Brooks, 2006: Lead time and time under tornado warnings: 1986 2004. Preprints, *23rd Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., 11.5. [Available online at https://ams.confex.com/ams/23SLS/techprogram/paper_115194.htm]

Farmer, E. W., A. J. Belyavin, A. J. Tattersall, A. Berry, and G. R. J. Hockey, 1991: Stress in air traffic control II: Effects of increased workload. RAF Institute of Aviation Medicine Report 701.

Farmer, E., and A. Brownson, 2003: Review of workload measurement, analysis and interpretation methods. Eurocontrol Agency Report CARE-Integra-TRS-130-02-WP2-1-0, 33 pp.

Fitts, P. M., R. E. Jones, and J. L. Milton, 1950. Eye movements of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review*, **9**, 24–29.

Flemisch F. O., and R. Onken, 2000: Detecting usability problems with eye tracking in airborne battle management support. *Proc. NATO RTO HFM Symposium on Usability of information in Battle Management Operations*. 2000, Oslo, Norway, RTO/NATO, 1–13.

Fletcher, J. O., 1990: Early developments of weather radar during World War II. *Radar in Meteorology: Battan Memorial and 40th Anniversary Radar Meteorology Conference*. D. Atlas, Ed., Amer. Meteor. Soc., 3–6.

Forsyth, D. E., and Coauthors, 2005: The National Weather Radar Testbed (phased array). Preprints, *32nd Conf. on Radar Meteorology*, Albuquerque, NM, Amer. Meteor. Soc., 12R.3. [Available online at https://ams.confex.com/ams/pdfpapers/96377.pdf.]

Friday, E., 1994: The modernization and associated restructuring of the National Weather Service: An overview. *Bull. Amer. Meteor. Soc.*, **75**, 43–52.

Friedman, J. R., C. Silva., H. Jenkins-Smith, and P. Spicer, 2015: Public, publics, and social media: Ethnographic observations on forecasters navigating uncertainty regarding social media. *10th Symposium on Societal Applications: Policy, Research, and Practice*, Phoenix, AZ, Amer. Meteor. Soc., 9.5. [Abstract and presentation online: https://ams.confex.com/ams/95Annual/webprogram/Paper269833.html]

Funk, T. W., K. E. Darmofal, J. D. Kirkpatrick, V. L. DeWald, R. W. Przybylinski, G. K. Schmocker, and Y. J. Lin, 1999: Storm reflectivity and mesocyclone evolution associated with the 15 April 1994 squall line over Kentucky and southern Indiana. *Wea. Forecasting*, **14**, 976–993.

Gallo, B. A., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed spring forecasting experiment. *Wea. Forecasting*, in press.

Gidlöf, K., A., Wallin, R. Dewhurst, K. Holmqvist, 2013: Using eye tracking to trace a cognitive process: Gaze behavior during decision making in a natural environment. *Journal of Eye Movement Research*, **6**, 1–14.

Giovinco, N. A., S. M. Sutton, J. D. Miller, T. M. Rankin, G. W. Gonzalez, B. Najafi, and D. Armstrong, 2015: A passing glance? Differences in eye tracking and gaze patterns between trainees and experts reading plain film bunion radiographs. *The Journal of Foot and Ankle Surgery*, **54**, 382–391.

Girard, J. M., M. Wilczyk, Y. Barloy, P. Simon, and J. C. Popieul, 2005: Towards an on line assessment of subjective driver workload. *Driving Simulation Conference North America*, Orlando, Florida, University of Iowa, 382–391.

Goldberg, J. H., M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky, 2002: Eye tracking in web search tasks: Design implications. Proc. *2002 symposium on Eye Tracking Research & Applications Symposium*. 2002, New Orleans, LA, New York ACM, 51–58.

Goldberg, H. J., and A. M. Wichansky, 2003: Eye Tracking in Usability Evaluation: A Practitioner's Guide. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, J. Hyönä, R. Radach, and H. Deubel, Eds., Amsterdam: Elsevier, 493–516.

Goodman, S. J., and Coauthors, 2012: The GOES-R Proving Ground: Accelerating user readiness for the next-generation geostationary environmental satellite system. *Bull. Amer. Meteor. Soc.*, **93**, 1029–1040.

Gore, J., R. Flin, N. Stanton, and B. L. W. Wong, 2015: Applications for naturalistic decision making. *Journal of Occupational and Organizational Psychology*, **88**, 223–230.

Guastello, S. J., A. Shircel, M. Malon, and P. Timm, 2015: Individual differences in the experience of cognitive workload. *Theoretical Issues in Ergonomics Science*, **16**, 20–52.

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

Hart, S. G., and L. E. Staveland, 1988: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, P. A. Hancock, and N. Meshkati, Eds., Amsterdam: Elsevier, 139–183.

Hart, S. G., 2006: NASA-task load index (NASA-TLX); 20 years later. *Proc. 50th Annual Meeting of the Human Factors and Ergonomics Society*, San Francisco, CA, Human Factors and Ergonomics Society, 904–908.

Hauland, G., 2008: Measuring individual and team situation awareness during planning tasks in training of en route air traffic control. *The International Journal of Aviation Psychology*, **18**, 290–304.

Hawkins, M. D., V. Brown, and J. Ferrell, 2017: Assessment of NOAA National Weather Service methods to warn for extreme heat events. *Wea. Forecasting*, **9**, 5–13.

Heinselman, P. L., D. L. Priegnitz, K. L. Manross, T. M. Smith, and R. W. Adams, 2008: Rapid sampling of severe storms by the National Weather Radar Testbed phased array radar. *Wea. Forecasting*, **23**, 808–824.

Heinselman, P. L. and S. M. Torres, 2011: High-temporal-resolution capabilities of the National Weather Radar Testbed phased-array radar. *J. Appl. Meteor. Climatol.*, **50**, 579–593.

Heinselman, P. L., D. S. LaDue, and H. Lazrus, 2012: Exploring impacts of rapid-scan radar data on NWS decisions. *Wea. Forecasting*, **27**, 1031–1044.

Heinselman, P., D. LaDue, D. M. Kingfield, and R. Hoffman, 2015: Tornado warning decisions using phased array radar data. *Wea. Forecasting*, **30**, 57–78.

Henderson, J. M. & Ferreira, F., 2004: *The Integration of Language, Vision, and Action: Eye Movements and the Visual World*. New York: Psychology Press.

Henderson, J., R. S. Schumacher, and E. R. Nielsen, 2017: Tornado and flash flood (TORFF) warnings: Operational challenges during multiple hazards. *Special Symposium on Severe Local Storms: Observation Needs to Advance Research, Prediction, and Communication*, Seattle, WA, Amer. Meteor. Soc., 2.3. [Abstract and presentation online: https://ams.confex.com/ams/97Annual/webprogram/Paper315198.html]

Hendrickson, J. J., 1989: Performance, preference, and visual scan patterns on a menu based system: Implications for interface design. *Proc. ACM CHI 89 Human Factors in Computing Systems Conference*. 1989, Austin, TX, ACM Press, 2317–222.

Hering, H., and G. Coatleven, 1996: ERGO (version 2) for instantaneous self assessment of workload in a real time ATC simulation environment. Eurocontrol Agency Report No. 10/96, 52 pp. [Available online at http://www.eurocontrol.int/sites/default/files/library/ 014_ERGO.pdf]

Hervet, G., K. Guérard, S. Tremblay, and M. S. Chtourou, 2011: Is banner blindness genuine? Eye tracking internet text advertising. *Applied Cognitive Psychology*, **25**, 708–716.

Hilburn, B., 2003: Evaluating human interaction with advanced air traffic management automation. National Aerospace Laboratory Netherlands Report RTO-MP-088, 12 pp. [Available online at http://www.dtic.mil/dtic/tr/fulltext/u2/a422128.pdf]

Hitschfield, W. F., 1986: The invention of radar meteorology. *Bull. Amer. Meteor. Soc.*, **67**, 33–37.

Hoffman, R. R., 2005: Protocols for cognitive task analysis. DTIC No. ADA475456, 108 pp.

Hoium, D. K., A. J. Riordan, J. Monahan, and K. K. Keeter, 1997: Severe thunderstorm and tornado warnings at Raleigh, North Carolina. *Bull. Amer. Meteor. Soc.*, **78**, 2559–2575.

Holmqvist, K., M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, 2011: *Eye Tracking: A Comprehensive Guide to Methods and Measures.* Oxford University Press, 537 pp.

Huey, E. B., 1908: *The psychology and pedagogy of reading*. New York: Macmillan. 469 pp

Hvelplund, K. T., 2014: Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. *MonTI Special Issue - Minding Translation*, R. M. Martin Ed., Publicaciones de la Universidad de Alicante. 201–224.

Isom, B., and Coauthors, 2013: The Atmospheric Imaging Radar: Simultaneous volumetric observations using a phased array weather radar. *J. Atmos. Oceanic Technol.*, **30**, 655–675.

Istok, M. J., and Coauthors, 2009: WSR-88D dual-polarization initial operational capabilities. Preprints, *25th Conf. on International Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, Amer. Meteor. Soc., 15.5.

Ivìc, I. R., 2017: An approach to simulate the effects of antenna patterns n polarimetric variable estimates. *JTECH*, in print.

Jacob, R. J., and K. S. Karn, 2003: Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, J. Hyönä, R. Radach, and H. Deubel, Eds., Elsevier Science: Amsterdarm, 573–605.

Jansen, R. J., B. D. Sawyer, R. Van Egmond, H. De Ridder, and P. A. Hancock, 2016: Hysteresis in mental workload and task performance: The influence of demand transitions and task prioritization. *Human Factors*, **58**, 1143–1157.

Jarodzka, H., K. Holmqvist, and M. Nyström, 2010: A vector-based, multidimensional scanpath similarity measure. *Proc. 2010 Symposium on Eye-Tracking Research and Applications*, New York, NY: ACM, 211–218.

Jarodzka, H., K. Holmqvist, and H. Gruber, 2017: Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research*, **10**, 1–18.

Jordan, C. S., 1992: Experimental study of the effect of an instantaneous self assessment workload recorder on task performance, Defence Research Agency Portsmouth UK Tech. Note DRA/TM/CAD5/92011.

Jorna. P., 1997: Human machine interfaces for ATM: Objective and subjective measurements on human interactions with future flight deck and air traffic control systems. Proceedings of the USA/FAA Air Traffic Management R&D Seminar 1997, Saclay, France. [Available online at http://www.atmseminarus.org/seminarContent/seminar1/papers/p_006_CDR.pdf]

Judd, C. H., C. N. McAllister, and W. M. Steel, 1905: *General Introduction to a Series of Studies of Eye Movements by Means of Kinetoscopic Photographs*. Baltimore: The Review Publishing Company, 1–16.

Just, M. A., and P. A. Carpenter, 1976a: Eye fixations and cognitive processes. *Cognitive Psychology*, **8**, 441–480.

Just, M. A., and P. A. Carpenter, 1976b: The role of eye-fixation research in cognitive psychology. *Behavior Research Methods & Instrumentation*, **8**, 139–143.

Just, M. A., and P. A. Carpenter, 1980: A theory of reading: From eye fixations to comprehension. *Psychological Review*, **87**, 329–354.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The spring program. *Bull. Amer. Meteor. Soc.*, **84**, 1797-1806.

Kang, Z., and S. J. Landry, 2014: Using scanpaths as a learning method for a conflict detection task of multiple target tracking. *Human Factors*, **56**, 1151–1162.

Karstens, C., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting,* **30**, 1551–1570.

Karstens, C. D., and CoAuthors, 2016: Forecaster Decision-Making with Automated Probabilistic Guidance in the 2015 Hazardous Weather Testbed Probabilistic Hazard Information Experiment. *Fourth Symposium on Building a Weather-Ready Nation*, New Orleans, LA, Amer. Meteor. Soc. [Abstract and presentation online: https://ams.confex.com/ams/96Annual/webprogram/ Paper286854.html]

Khan, R. S. A., G. Tien. M. S. Atkins, B. Zheng, O. N. M. Panton, and A. T. Meneghetti, 2012: Analysis of eye gaze: Do novice surgeons look at the same location as expert surgeons during a laparoscopic operation? *Surgical Endoscopy*, **26**, 3536–3540.

Kirwan, B., A. Evans, L. Donohoe, A. Kilner, T. Lamoureux, T. Atkinson, and H. MacKendrick, 1997: Human factors in the ATM system design life cycle. FAA/Eurocontrol ATM R&D Seminar, Paris, France. Available online at [http://www.atmseminar.org/seminarContent/seminar1/papers/p_007_CDR.pdf].

Klein, G., 2008: Naturalistic decision making. *Human Factors*, **50**, 456–460.

Kruger, R. A., and M. A. Casey, 2015: *Focus Groups: A Practical Guide for Applied Research*. SAGE Publications Inc., 280 pp.

Kruskall, W. H., and W. A. Wallis, 1952: Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47** 583–621.

Kumjian, M. R., and A. V. Ryzhkov, 2008: Polarimetric signatures in supercell thunderstorms. *J. Appl. Meteor. Climatol.*, **48**, 1940–1961.

Kumjian, M. R., 2013: Principles and applications of dual-polarization weather radar. Part I: Description of the polarimetric radar variables. *J. Operational Meteor*., **1**, 226–242.

Kurdzo, J. M., D. J. Bodine, B. L. Cheong, and R. D. Palmer, 2015: High-temporal resolution polarimetric X-band Doppler radar observations of the 20 May 2013 Moore, Oklahoma, tornado. *Mon. Wea. Rev.*, **143**, 2711–2735.

Kuster, C. M., P. L. Heinselman, and M. Austin, 2015: 31 May 2013 El Reno tornadoes: Advantages of rapid-scan phased-array radar data from a warning forecaster's perspective. *Wea. Forecasting*, **30**, 933–956.

Kuster, C. M., P. L. Heinselman, and T. J. Schuur, 2016: Rapid-update radar observations of downbursts occurring within an intense multicell thunderstorm on 14 June 2011. *Wea. Forecasting*, **31**, 824–851.

Kuster, C. M., J. C. Snyder, P. L. Heinselman, and T. J. Schuur, 2017a: Rapid-scan dual-polarization radar observations of ZDR column depth in the context of forecaster conceptual models. *38th Conf. on Radar Meteorology*, Chicago, IL, Amer. Meteor. Soc., 19A.5. [Available online at https://ams.confex.com/ams/38RADAR/meetingapp.cgi/ Paper/320591].

Kuster, C. M., P. L. Heinselman, J. C. Snyder, K. A. Wilson, D. A. Speheger, and J. E. Hocker, 2017b: Bulding community with public safety officials in Oklahoma: Evaluating use of radar-based tornado track estimation products. *Wea. Forecasting*, in press.

LaDue, D., P. Heinselman, and J. Newman, 2010: Strengths and limitations of current radar systems for two stakeholder groups in the southern plains. Bull. Amer. Meteor. Soc., **91**, 899–910.

LaDue, D., and Coauthors, 2017: Temporal and Spatial Aspects of Emergency Manager Use of Prototype Probabilistic Hazard Information. *Fifth Symposium on Building a Weather-Ready Nation: Enhancing Our Nation's Readiness, Responsiveness, and Resilience to High Impact Weather Events*, Seattle WA, Amer. Meteor. Soc., 312923.

Lakshmanan, V., T. Smith, G. J. Stumpf, and K. Hondl: 2007: The warning decision support system - integrated information. *Wea. Forecasting*, **22**, 596–612.

Lazar, J., J. H. Feng, and H. Hochheiser, 2010: *Research Methods in Human-Computer Interaction*. John Wiley and Sons Ltd, 426 pp.

Lemercier, C., and Coauthors, 2014: Inattention behind the wheel: How factual internal thoughts impact attentional control while driving. *Safety Science*, **62**, 279–285.

Lemon, L. R., R. J. Donaldson, D. W. Burgess, and R. A. Brown, 1977: Doppler radar application to severe thunderstorm study and potential real-time warning. *Bull. Amer. Meteor. Soc.*, **58**, 1187–1193.

Levenshtein, V., 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics–Doklady*, **10**, 707–710.

Lindell, M. K., and H. Brooks, 2013: Workshop on Weather Ready Nation: Science imperatives for severe thunderstorm research. *Bull. Amer. Meteor. Soc.*, **94**, 171–174.

Line, W. E., 2014: GOES-R Proving Ground demonstration at the Hazardous Weather Testbed 2014 Spring Experiment final evaluation. NOAA/Hazardous Weather Testbed Rep., 36 pp. [Available online at http://www.goes-r.gov/users/docs/pg-activities/PGFR-HWT-2014-Final.pdf.]

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. *Proc. Human Factors Ergonomics Society Annual Meeting*, **59**, 1791–1795.

Ling, C., and Coauthors, 2017: Forecasters' mental workload while issuing probabilistic hazard information (PHI) during 2016 FACETs PHI Hazardous Weather Testbeds. *33rd Conf. on Environmental Information Processing Technologies*, Seattle, WA, Amer. Meteor. Soc., J9.3. [Available online https://ams.confex.com/ams/97Annual/webprogram/Paper314172.html]

Lipshitz, R., G. Klein, and J. Orasanu, 2001: Focus article: Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, **14**, 331–352.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Science*, **4**, 6–14.

Luximon, A., and R. S. Goonetilleke, 2001: Simplified subjective workload assessment technique. *Ergonomics*, **44**, 229–243.

Mackworth, J. F., and N. H. Mackworth, 1958: Eye fixations recorded on changing visual scenes by the television eye-marker. *Journal of the Optical Society of America*, **48**, 439–445.

Manning, D. J., S. C. Ethell, and T. Donovan, 2014: Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *The British Journal of Radiology*, **77**, 231–235.

Marquart, G., C. Cabrall, and J. De Winter, 2015: Review of eye-related measures of drivers' mental workload. *6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences*, Las Vegas, NV, Applied Human Factors and Ergonomics, **3**, 2854–2861.

Matin, E., 1974: Saccadic suppression: A review and an analysis. *Psychological Bulletin*, **81**, 899-917.

McLaughlin, D., and Coauthors, 2009: Short-wavelength technology and the potential for distributed networks of small radar systems. *Bull. Amer. Meteor. Soc.*, **90**, 1797–1817.

Mehta, R. K., and R. Parasuraman, 2013: Neuroergonomics: A review of applications to physical and cognitive work. *Frontiers in Human Neuroscience*, **7**, 889.

Merton, R. K., 1968: The Matthew effect in science. *Science*, **159**, 56–63.

Metcalf, J. I., and K. M. Glover, 1990: A history of weather radar research in the U.S. Air Force. *Radar in Meteorology: Battan Memorial and 40th Anniversary Radar Meteorology Conference*. D. Atlas, Ed., Amer. Meteor. Soc., 32–41.

Miller, S., 2001: Literature review: Workload Measures. National Advanced Diving Simulator Report N01-006, 65 pp. [Available online at http://www.nads-sc.uiowa.edu/publicationstorage/200501251347060.n01-006.pdf]

Morss, R. E., O. V. Wilhelmi, M. W. Downton, and E. Gruntfest, 2005: Flood risk, uncertainty, and scientific information for decision making: Lessons from an interdisciplinary project. Bull. Amer. Meteor. Soc., 90, 1797–1817.

Morss, R. E., and F. M. Ralph, 2007: Use of information by National Weather Service forecasters and emergency managers during CALJET and PACJET-2001. *Wea. Forecasting*, **22**, 539–555.

Murphy, A. H., and R. L. Winkler, 1971: Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.*, **52**, 239–248.

Murphy, A. H., and R. L. Winkler, 1974: Probability forecasts: A survey of National Weather Service forecasters. *Bull. Amer. Meteor. Soc.*, **55**, 1449–1452.

National Academies of Science, 2012: Weather services for the nation: Becoming second to none, National Academies Press, 86 pp.

NCEI, 2016: Storm Events Database. Accessed 7 November 2016 [Available online at https://www.ncdc.noaa.gov/stormevents/]

Neumann, D. L., and O. V. Lipp, 2002: Spontaneous and reflexive eye activity measures of mental workload. *Australian Journal of Psychology*, **54**, 174–179.

Newman, J. F., and P. L. Heinselman, 2012: Evolution of a quasi-linear convective system sampled by phased array radar. *Mon. Wea. Rev.*, **140**, 3467–3486.

NOAA, 2015: Weather Ready Nation. Accessed 25 November 2015. [Available online at http://www.nws.noaa.gov/com/weatherreadynation/#.Vk5N5vlVhBd]

NOAA, 2017a: About NOAA's National Weather Service. Accessed 14 March 2017. [Available online at http://www.weather.gov/about]

NOAA, 2017b: Weather Ready Nation. Accessed 14 March 2017. [Available online at http://www.nws.noaa.gov/com/weatherreadynation/#.Vk5N5vlVhBd]

Noton, D., and L. Stark, 1971: Scanpaths in saccadic eye movement while viewing and recognizing patterns. *Vision Research*, **11**, 929–942.

Nourbakhsh, N., W. Wang, F. Chen, and R. A. Calvo, 2012: Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. *Proc. 24th Australian Computer-Human Interaction Conference*. Melbourne, Victoria, Association for Computing Machinery, 420–423.

Novak, D. R., D. R. Bright, and M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069–1084.

NWS, 2015: Verification. NWS Instruction 10-1601. [Available online at http://www.nws.noaa.gov/directives]

Obermeier, H., K. L. Nemunaitis-Berry, S. A. Jasko, D. LaDue, C. Karstens, G. M. Eosco, A. Gerard, and L. Rothfusz, 2017: Broadcast Meteorologist Decision Making in the 2016 Hazardous Weather Testbed Probabilistic Hazard Information Project, *12th Symposium on Societal Applications: Policy, Research and Practice*, Seattle WA, Amer. Meteor. Soc., 4.3.

Olsen, M., 2012: The Tobii I-VT fixation filter: Algorithm description. Accessed 12 November 2015. [Available online at https://stemedhub.org/resources/2173/download/Tobii_WhitePaper_TobiiIVTFixationFilter.pdf]

Parasuraman, R., 2011: Neuroergonomics: Brain, cognition, and performance at work. *Current Directions in Psychological Science*, **20**, 181–186.

Pazmany, A. L., J. B. Mead, H. B. Bluestein, J. C. Snyder, and J. B. Houser, 2013: A mobile rapid-scanning x-band polarimetric (RaXPol) Doppler radar system. *J. Atmos. Oceanic Technol.*, **30**, 1398–1413.

Poole, A., and L. J. Ball, 2006: Eye tracking in human-computer interaction and usability research: Current status and future. *Encyclopedia of human-computer interaction.* C. Ghaoui, Ed., Pennsylvania: Idea Group Inc., 211–219.

Proctor, R. W., and T. Van Zandt, 2008: Human Factors in Simple and Complex Systems. CRC Press, 696 pp.

Qang, Q., S. Yang, M. Liu, Z. Cao, and Q. Ma, 2014: An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems*, **62**, 1–10.

Quoetone, E., J. Boettcher, and C. Spannagle, 2009: How did that happen? A look at factors that go into forecaster warning decisions. Extended Abstracts, *34th National Weather Association Annual Meeting*, Norfolk, VA. [Available online at www.nwas.org /meetings/nwa2009/.]

Ralph, F. M. and Coauthors, 2013: The emergence of weather-related test beds linking research and forecasting operations. *Bull. Amer. Meteor. Soc.*, **94**, 1187–1211.

Rasmussen, E., 2015: VORTEX-Southeast program overview. National Severe Storms Laboratory Tech. Report, 36pp. [Available online at ftp://ftp.atdd.noaa.gov/pub/ vortexse/ProjectOverview.pdf]

Rayner, K., 1998: Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**, 372–422.

Rayner, K., A. Pollatsek, J. Ashby, and C. Clifton Jr., 2012: *Psychology of Reading: Second Edition*. Psychology Press, 485 pp.

Reid, G. B., C. A. Shingledecker, and F. T. Eggemeier, 1981: Application of conjoint measurement to workload scale development. *Proc. of the Human Factors Society 25th Annual Meeting.* Rochester, NY, Human Factors Society, 522–526.

Reid, G. B, and T. E. Nygren, 1988: The subjective workload assessment technique: A scaling procedure for measuring mental workload. Human Mental Workload, P. A. Hancock, and N. Meshkati, Eds., Amsterdam: Elsevier, 185–218.

Resnick, M. L., and W. Albert, 2013: The impact of advertising location and user task on the emergence of banner ad blindness: An eye tracking study. *Proc. Human Factors and Ergonomics Society Annual Meeting*. 2013, San Diego, CA, Human Factors and Ergonomics Society, 1037–1041.

Roberts, R. D., and J. W. Wilson, 1989: A proposed microburst nowcasting procedure using single-Doppler radar. *J. Appl. Meteor.*, **28**, 285–303.

Roscoe, A. H., 1992: Assessing pilot workload. Why measure heart rate, HRV, and respiration? *Biological Psychology*, **34**, 259–288.

Rose, C. L., L. B. Murphy, L. Byard, and K. Nikzad, 2002: The role of the big five personality factors in vigilance performance and workload. *European Journal of Personality*, **16**, 185–200.

Rothfusz, L. P., C. Karstens, and D. Hilderband, 2014: Next-generation severe weather forecasting and communication. *Eos*, **95**, 325–326.

Ryzhkov, A. V., T. J. Schuur, D. W. Burgess, P. L. Heinselman, S. E. Giangrande, and D. S. Zrnić, 2005: The Joint Polarization Experiment: Polarimetric rainfall measurements and hydrometeor classification. *Bull. Amer. Meteor. Soc.*, **86**, 809–824.

Saffle, R. E., M. J. Istok, and G. Cate, 2009: NEXRAD product improvement – update 2009. *25th Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* Phoenix, AZ, Amer. Meteor. Soc., 10B.1.

Schapp, T. W., A. R. A. Van der Horst, B. Van Arem, and K. A. Brookhuis, 2013: The relationship between driver distraction and mental workload. *Driver Distraction and Inattention: Advances in Research and Countermeasures*, M. A. Regan, J. D. Lee, and T. W. Victor, Eds., Farnham: Ashgate, 63–80.

Scharfenberg, K. A., and Coauthors, 2005: The Joint Polarization Experiment: Polarimetric radar in forecasting and warning decision-making. *Wea. Forecasting*, **20**, 775–788.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI on the GOES-R series. *Bull. Amer. Meteor. Soc.*, **98**, 681–698.

Schuur, T., P. Heinselman, K. Scharfenberg, A. Ryzhkov, D. Zrnić, V. Melnikov, and J. Krause, 2003: Overview of the Joint Polarization Experiment (JPOLE). National Severe Storms Laboratory Report, 39pp. [Available online at http://www.nssl.noaa.gov/publications/wsr88d_reports/JPOLE_Overview_Report.pdf]

Schvartzman, D., S. Torres, and T. Yu, 2017: Weather radar spatiotemporal saliency: A first look at an information theory-based human attention model adapted to reflectivity images. *J. Atmos. Oceanic Technol.*, **34**, 137–152.

Sherman-Morris, K., K. B. Antonelli, and C. C. Williams, 2015: Measuring the effectiveness of the graphical communication of hurricane storm surge threat. *Wea. Climate Soc.*, **7**, 69–82.

Smith, R. L., and D. W. Holmes, 1961: Use of Doppler radar in meteorological observations. *Mon. Wea. Rev.*, **89**, 1–7.

Smith, T. M., and Coauthors, 2014: Examination of a real-time 3DVAR analysis system in the Hazardous Weather Testbed. *Wea. Forecasting*, **29**, 63–77.

Smith, T. M., and Coauthors, 2016: Multi-radar multi-sensor (MRMS) severe weather and aviation products. *Bull. Amer. Meteor. Soc.*, **97**, 1617-1630.

Stailey, J., and K. Hondl, 2016: Multifunction phased array radar for aircraft and weather surveillance, Proc. IEEE, **103**, 649–659.

Steadham, R., 2008: 2008 NWS field study. Part 1: Volume coverage pattern usage. Radar Operations Center, Norman, OK, 28 pp. [Available from WSR-88D Radar Operations Center, 1200 Westheimer Dr., Norman, OK 73069]

Stensrud, D., and Coauthors, 2009: Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499.

Sterman, M. B., and C. A. Mann, 1995: Concepts and applications of EEG analysis in aviation performance evaluation. *Biological Psychology*, **40**, 115–130.

Stumpf, G. J., T. M. Smith, K. Manross, and D. L. Andra, 2008: The Experimental Warning Program 2008 Spring Experiment at the NOAA Hazardous Weather Testbed. Preprints, *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 8A.1. [Available online at http://ams.confex.com/ams/pdfpapers/141712.pdf]

Sullivan, J., J. H. Yang, M. Day, and Q. Kennedy, 2011: Training simulation for helicopter navigation by characterizing visual scan patterns. *Aviation, Space, and Environmental Medicine*, **82**, 871–878.

Supinie, T. A., N. Yussouf, Y. Jung, M. Xue, J. Cheng, S. Wang, 2017: Comparison of the analyses and forecasts of a tornadic supercell storm from assimilating phased-array radar and WSR-88D observations. *Wea. Forecasting*, **32**, 1379–1401.

Svensson, E., M. Angelborg-Thanderz, L. Sjöeberg, and S. Olsson, 1997: Information complexity: Mental workload and performance in combat aircraft. *Ergonomics*, **40**, 362–380.

Szalma, J. L., 2002: Individual differences in the stress and workload of sustained attention. *Proc. Human Factors and Ergonomics Society 46th Annual Meeting*, **46**, 1002–1006.

Tanamachi, R. L., P. L. Heinselman, and L. J. Wicker, 2015: Impacts of a storm merger on the 24 May 2011 El Reno, Oklahoma, Tornadic Supercell. *Wea. Forecasting*, **30**, 501–524.

Tattersall, A. J., and P. S. Foord, 1996: An experimental evaluation of instantaneous self assessment as a measure of workload. *Ergonomics*, **39**, 740–748.

Tien, G., M. S. Atkins, B. Zheng, and C. Swindells, 2010: Measuring situation awarenessof surgeons in laparoscopic training. *Proc. 2010 Symposium on Eye-Tracking Research and Applications*. 2010, Austin, TX, ACM New York, 149–152.

Tien, T., P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, H. Yang, and A. Darzi, 2015: Differences in gaze behavior of expert and junior surgeons performing open inguinal hernia repair. *Surgical Endoscopy*, **29**, 405–413.

Tobii, 2014: User manual – Tobii Studio Version 3.3.0. Accessed May 2017. [Available online https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-tx300-eye-tracker-user-manual.pdf]

Tobii, 2017: Tobii Pro Research Solutions. Accessed May 2017 [Available online at https://www.tobiipro.com/]

Torres, S. M., and Coauthors, 2012: ADAPTS Implementation: Can we exploit phased array radar's electronic beam steering capabilities to reduce update time? Extended Abstract, *28th Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, New Orleans, LA, Amer. Meteor. Soc., 6B.3.

Trafton, J. G., S. Marshall, F. Mintz, and S. B. Trickett, 2002: Extracting explicit and implicit information from complex visualizations. *Diagrammatic Representation and Inference,* M. Heagarty, B. Meyer, and N. H. Narayanan, Eds., Springer-Verlag, 206–220.

Trapp, R. J., S. A. Tessendorf, E. S. Godfrey, and H. E. Brooks, 2005: Tornadoes from squall lines and bow echoes. Part I: Climatological distribution. *Wea. Forecasting* **20**, 23–34.

Van De Merwe, K., H. Van Dijk, and R. Zon, 2012: Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology*, **22**, 78–95.

Van Orden, K. F., W. Limbert, S. Makeig, and T. P. Jung, 2001: Eye activity correlates of workload during a visuospatial memory task. *Human Factors,* **43**, 111–121.

Vasiloff, S. V., 2001: Improving tornado warnings with the Federal Aviation Administration's Terminal Doppler Weather Radar. *Bull. Amer. Meteor. Soc.*, **82**, 861–874.

Veltman, J. A., and A. W. K. Gaillard, 1998: Physiological workload reactions to increasing levels of task difficulty. *Ergonomics,* **41**, 656–669.

Vidulich, M. A., and C. D. Wickens, 1986: Causes of dissociation between subjective workload measures and performance: Caveats for the use of subjective assessments. *Applied Ergonomics*, **17**, 291–296.

Vogt, J., T. Hagemann, and M. Kastner, 2006: The impact of workload on heart rate and blood pressure in en-route and tower air traffic control. *Journal of Psychophysiology*, **20**, 297–314.

Weber, M. E., J. Y. N. Cho, J. S. Herd, J. M. Flavin, W. E. Benner, and G. S. Torok, 2007: The next-generation multimission U.S. surveillance radar network. *Bull. Amer. Meteor. Soc.*, **88**, 1739–1751.

Whiton, R. C., P. L. Smith, S. G. Bigler, K. E. Wilk, and A. C. Harbuck, 1998a: History of operational use of weather radar by U.S. weather services, Part I: The pre-NEXRAD era. *Wea. Forecasting*, **13**, 219–243.

Whiton, R. C., P. L. Smith, S. G. Bigler, K. E. Wilk, and A. C. Harbuck, 1998b: History of operational use of weather radar by U.S. weather services. Part II: Development of operational Doppler weather radars. *Wea. Forecasting*, **13**, 244–252.

Wickens, C. D., and J. S. McCarley, 2008: *Applied Attention Theory*. London: CRC Press, 248 pp.

Wierwille, W. W., and J. G. Casali, 1983: A validated rating scale for global mental workload measurement applications. *Proc. Human Factors Society*, **27**, 129–133.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences, 2nd ed*. Academic Press, 467 pp.

Wilson, G. D., and F. T. Eggemeier, 1991: Physiological measures of workload in multi task environments. *Multiple-task performance,* D. F. Damos, Ed., London: Taylor & Francis, 329–360.

Wilson, G. F., 2002: An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, **12**, 3–18.

Wilson, J., R. Carbone, H. Baynton, and R. Serafin, 1980: Operational application of meteorological Doppler radar. *Bull. Amer. Meteor. Soc.*, **61**, 1154–1167.

Wilson, K. A., P. L. Heinselman, and Z. Kang, 2016: Exploring applications of eye tracking in operational meteorology research. *Bull. Amer. Meteor. Soc.*, **97**, 2019–2025.

Wilson, K. A., P. L. Heinselman, C. M. Kuster, and D. M. Kingfield, 2017: Forecaster performance and workload: Does radar update time matter? *Wea. Forecasting*, **32**, 253–274.

Winter, J. C. F., 2014: Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology & Work*, **16**, 289–297.

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

Wood, G., K. M. Knapp, B. Rock, C. Cousens, C. Roobottom, and M. R. Wilson, 2013: Visual expertise in detecting and diagnosing skeletal fractures. *Skeletal Radiology*, **42**, 165–172.

Wurman, J., D. Dowell, Y. Richardson, P. Markowski, E. Rasmussen, D. Burgess, L. Wicker, and H. B. Bluestein, 2012: The second Verification of the Origins of Rotation in Tornadoes Experiment: VORTEX2. *Bull. Amer. Meteor. Soc.*, **93**, 1147–1170.

Xue, M. Tong, and K. K. Droegemeier, 2006: An OSSE framework based on the ensemble square-root Kalman filter for evaluating impact of data from radar networks on thunderstorm analysis and forecast. *J. Atmos. Oceanic Technol.*, **23**, 46–66.

Yarbus, A. L, 1967: *Eye Movements and Vision*. Plenum Press, New York. 222 pp.

Yeh, Y. Y., and C. D. Wickens, 1988: Dissociation of performance and subjective measures of workload. *Human Factors*, **30**, 111–120.

Young, M. S., and N. A. Stanton, 2002: Mental workload: Theory, Measurement, and Application. *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Boca Raton, FL: CRC Press, 818–812.

Young, M. S., N. A. Stanton, 2005: Mental Workload. *Handbook of Human Factors and Ergonomics Methods*, N. A. Stanton, E. Salas, H. W. Hendrick, A. Hedge, and K. Brookhuis, Eds., CRC Press LLC, 39-1–39-7.

Young, M. S., K. A. Brookhuis, C. D. Wickens, and P. A Hancock, 2015: State of science: Mental workload in ergonomics. *Ergonomics*, **58**, 1–17.

Yu, C. S., E. E. M. Wang, W. C. Li, G. Braithwaite, N. Greaves, 2016: Pilots' visual scan patterns and attention distribution during the pursuit of a dynamic target. *Aerospace Medicine and Human Performance*, **87**, 40–47.

Yussouf, N., and D. J. Stensrud, 2010: Impact of phased-array radar observations over a short assimilation period: Observing system simulation experiments using an ensemble Kalman filter. *Mon. Wea. Rev.*, **138**, 517–538.

Zheng, B., G. Tien, S. M. Atkins, C. Swindells, H. Tanin, A. Menegetti, K. A. Qayumi, and O. N. M. Panton, 2011: Surgeon's vigilance in the operating room. *The American Journal of Surgery*, **201**, 846–848.

Zrnić, D. S., and A. V. Ryzhkov, 1999: Polarimetry for weather surveillance radars. *Bull. Amer. Meteor. Soc.*, **80**, 389–406.

Zrnić, D. S., and Coauthors, 2007: Agile-beam phased array radar for weather observations. *Bull. Amer. Meteor. Soc.*, **88**, 1753–1766.

Appendix A

Focus Group Questions

1. What was your first reaction to the 1-min, 2-min, and 5-min PAR update times?

2. How did your reactions to the 1-min, 2-min, and 5-min update times impact your interrogation strategies when working what you believed to be a a) severe hail and wind event, b) tornado event, and c) non-severe event?

3. Did you have a difference in understanding of what you believed to be a a) hail and wind event, b) tornado event, and c) non-severe event based on the temporal resolution of PAR data available?

4. Imagine you are going back to your office and you have rapid-update PAR data (1-min or 2-min updates) like you had here. Based on your 2015 PARISE experience, what concerns do you specifically have about using rapid-update PAR data in an operational sense?

5. Drawing from your 2015 PARISE experience, what kind of training do you think you would find useful in transitioning rapid-update PAR data into operations?

6. Imagine you are going back to your office and you have rapid-update PAR data (1-min or 2-min updates) like you had here. Based on the 2015 PARISE experience, how do you envision these radar data being integrated into your fuller warning decision process where you have your normal available data and are working with your colleagues?

7. What other thoughts or ideas from the week would you like to share with us?

Appendix B

List of Acronyms

ADAPTS	Adaptive Digital Signal Processing Algorithm for PAR Timely Scans
AOI	Area of Interest
AVSET	Automated Volume Scan Evaluation and Termination
FACETS	Forecasting a Continuum of Environmental Threats
FAR	False Alarm Ratio
GOES-R	Geostationary Operational Environmental Satellite R
ISA	Instantaneous Self-Assessment
JDOP	Joint Doppler Operation Project
JPOLE	Joint Polarization Experiment
MESO-SAILS	Multiple Elevation Scan Option—Supplemental Adaptive Intravolume Low-Level Scan
NASA-TLX	NASA-Task Load Index
NOAA	National Oceanic and Atmospheric Administration
NWS	National Weather Service
PAR	Phased-Array Radar
PARISE	Phased Array Radar Innovative Sensing Experiment
POD	Probability of Detection
SAILS	Supplemental Adaptive Intravolume Low-Level Scan
WDSS-II	Warning Decision Support System-Integrated Information
WFO	Weather Forecast Office
WSR-88D	Weather Surveillance Radar 1998 Doppler