UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SUPPORTING SITUATION AWARENESS AND DECISION MAKING IN WEATHER FORECASTING

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ELIZABETH MINTMIRE ARGYLE
Norman, Oklahoma
2016

SUPPORTING SITUATION AWARENESS AND DECISION MAKING
IN WEATHER FORECASTING

A DISSERTATION APPROVED FOR THE
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY

____
Dr. Randa L. Shehab, Co-Chair

____
Dr. Ziho Kang, Co-Chair

____
Dr. Jonathan J. Gourley

____
Dr. Scott Gronlund

____
Dr. Pamela Heinselman

____
Dr. Theodore Trafalis

Acknowledgments

This work would not be complete without expressing gratitude to all of my friends, family, and colleagues who have supported me throughout this work. First, I would like to thank my co-advisors, Dr. Randa L. Shehab and Dr. Ziho Kang, for their support. Since my undergraduate days, Dr. Shehab has inspired me to aim high, to work hard, and to be persistent, leading me to where I am now. In addition, I greatly appreciate Dr. Kang's feedback and insights. Many of the ideas and methods contained within this dissertation matured as a result of numerous productive discussions with both Dr. Shehab and Dr. Kang.

I would also like to thank Dr. Jonathan Gourley and the members of the Flooded Locations and Simulated Hydrographs (FLASH) team, who gave me many first-hand opportunities to learn about the world of weather forecasting. As a member of the team, I have grown to appreciate and enjoy conducting interdisciplinary research. I have valued working with and learning from Zac, Race, Jess, Jill, Galateia, Steve, Ami, Humberto, and Pierre. Each of them provided valuable feedback, time, and expertise during this research, for which I am especially grateful.

I would not be where I am today were it not for Dr. Chen Ling. In the years we have worked together, she has been a role model, mentor, and friend. I would like to express my utmost thanks to her for introducing me to the field of human factors, for fostering my interest in research, and for encouraging me to pursue new opportunities.

My greatest thanks must be given to my family and friends. Throughout my life, my parents and sister have instilled values into me, particularly in relation to education, exploration, and creativity. I am immensely grateful for their advice,

encouragement, and support that has shaped who I am today. In addition, I wish to thank my good friends, Rachel and Natalie, whose friendship has been a foundation for me over many years. Last but certainly not least, I would like to thank Peter for his patience, empathy, and support over the years. I am truly thankful to have had such a friend and partner supporting me in this endeavor.

Table of Contents

List of Tables

List of Figures

Abstract

Weather forecasting is full of uncertainty, and as in domains such as air traffic control or medical decision making, decision support systems can affect a forecaster's ability to make accurate and timely judgments. Well-designed decision aids can help forecasters build situation awareness (SA), a construct regarded as a component of decision making. SA involves the ability to perceive elements within a system, comprehend their significance, and project their meaning into the future in order to make a decision. However, how SA is affected by uncertainty within a system has received little attention. This tension between managing uncertainty, situation assessment, and the impact that technology has on the two, is the focus of this dissertation.

To address this tension, this dissertation is centered on the evaluation of a set of coupled models that integrate rainfall observations and hydrologic simulations, coined "the FLASH system" (Flooded Locations and Simulated Hydrographs project). Prediction of flash flooding is unique from forecasting other weather-related threats due to its multi-disciplinary nature. In the United States, some weather forecasters have limited hydrologic forecasting experience. Unlike FLASH, current flash flood forecasting tools are based upon rainfall rates, and with the recent expansion into coupled rainfall and hydrologic models, forecasters have to learn quickly how to incorporate these new data sources into their work. New models may help forecasters to increase their prediction skill, but no matter how far the technology advances, forecasters must be able to accept and integrate the new tools into their work in order to gain any benefit. A focus on human factors principles in the design stage can help to ensure that by the time the product is transitioned into operational use, the decision support system addresses users' needs while minimizing task time, workload, and attention constraints.

This dissertation discusses three qualitative and quantitative studies designed to explore the relationship between flash flood forecasting, decision aid design, and SA. The first study assessed the effects of visual data aggregation methods on perception and comprehension of a flash flood threat. Next, a mixed methods approach described how forecasters acquire SA and mitigate situational uncertainty during real-time forecasting operations. Lastly, the third study used eye tracking assessment to identify the effects of an automated forecasting decision support tool on SA and information scanning behavior. Findings revealed that uncertainty management in forecasting involves individual, team, and organizational processes. We make several recommendations for future decision support systems to promote SA and performance in the weather forecasting domain.

Chapter 1: Introduction

On the morning of 10 June 2014, a major flash flood swept through Prince George's County, Maryland. With little warning, local residents found themselves amidst house flooding while drivers became stranded in their vehicles. Emergency management services reported eleven incidents including high water rescues from vehicles and evacuations of stranded homeowners from flooded buildings (National Climatic Data Center, 2014). The Washington Post reported that at least twenty-four rescues occurred and that some local residents evacuated to an emergency shelter in a local school (Bui, 2014).

While the local National Weather Service (NWS) Weather Forecast Office (WFO) had issued a flash flood warning at 9:28 AM EDT (National Weather Service, 2014a), some residents were not able to take necessary precautions in advance of the flooding (Halverson, 2014). In the warning text issued at 9:28 AM, the Sterling WFO wrote: "At 9:24 AM EDT… National Weather Service Doppler radar indicated very heavy rain capable of producing flash flooding. Additional rainfall amounts of 1 to 2 inches can be expected" (National Weather Service, 2014a). However, according to local media, the reality was that the area received up to five inches of rain in just two hours (Halverson, 2014). The Baltimore Sun reported that the rainfall stopped around 11:00 AM EDT, allowing the floodwaters to recede (Rector, 2014). Despite the issued warning, some considered this to be a "missed" event due to the short lead time given to locals (Halverson, 2014).

Could anything have provided more lead time to those affected by this event? Although forecasters had relevant training as well as access to computational models and observational tools, the unfolding rainfall event showed minimal chance of producing flash floods (Halverson, 2014). In this case, Halverson (2014) posited that false expectations were in part due to a lack of high-resolution gridded flash flooding and rainfall prediction models and few observational data sets. In order to assist users in drawing connections between conceptual models and environmental dynamics, some researchers have called for the development of analysis tools, referred to in the current work as forecasting decision support systems (Stuart et al., 2006; Trafton & Hoffman, 2007).

Decision support systems are information technology products that aid users in making efficient and effective decisions (Shim et al., 2002). Advances to forecasting decision support systems may improve outcomes if systems complement the way in which forecasters create, update, and implement their mental models (Trafton & Hoffman, 2007). One promising line of research involves the development of decision support tools that automate parts of the situation assessment process. Automation is frequently used to reduce workload and time pressures, allowing the operator to allocate his or her attention to other aspects of the work (Röttger, Bali, & Manzey, 2009). Furthermore, decision support systems are viewed as a low level of automation, in which the system provides guidance to a user who is in control of the decision and resulting action (Endsley & Kiris, 1995). In the weather forecasting domain, an appropriate level of decision support may promote situation awareness development, which could help to reduce missed weather events.

Situation awareness (SA) is regarded as an integral component of the decision making process involved in professional forecasting (Quoetone, Andra, Bunting, &

Jones, 2001). Although sometimes referred to as “situational awareness” in operational settings (Byrne, 2015), here, we will adopt the term most frequently used in theoretical research (“situation awareness”) as supported by Endsley’s 1995 Model of SA (Endsley, 1995c).

SA is a measurable construct and reflects an individual’s degree of knowledge regarding the state of their environment (Endsley, 1995a, 1995c). In one of the most widely-accepted models of SA, Endsley (1995c) defined SA as a construct with three levels. Level 1 SA comprises an individual’s perception of the environment, while Level 2 SA involves comprehension, or turning the perceived information into meaning. Level 3 SA (projection) centers around an individual’s ability to project the current state of the environment correctly into a likely future state. SA is not a static construct, but updates over time as decision makers gain experience with similar situations. Additional mechanisms such as information processing, memory, goals, preconceptions, background training, and system design also contribute to building and maintaining high levels of SA. In the current work, we follow the precedent set by Endsley (1995c, 2015b) and distinguish the measurable product (situation awareness; SA), from the process in which SA is developed and maintained (situation assessment).

Grounded in the field of human factors, this work explores the role of decision support system design on the situation assessment process in the weather forecasting domain. Throughout this dissertation, we investigate SA in weather forecasting from a qualitative and a quantitative standpoint; in doing so, we are able to identify behavioral patterns that facilitate accurate situation assessment while also developing recommendations for decision support system design. Although some studies have described the situation assessment process for experts in fields like air traffic control (Dao et al., 2009; Moore & Gugerty, 2010; van de Merwe, Oprins, Eriksson, & van der Plaat, 2012) and driving (Endsley & Kiris, 1995; Moore, 2009), a smaller number have provided empirical support for the situation assessment process in weather forecasting (Bowden & Heinselman, 2016; Jones, Quoetone, Ferree, Magsig, & Bunting, 2003; Quoetone et al., 2001). Understanding how forecasters develop SA will lead to improvements in forecast lead time and accuracy if we can find new ways to convey information to forecasters, particularly in heavy-workload, time-sensitive forecast situations.

In addition to extending theoretical accounts of SA to the weather forecasting domain, this work is motivated by the impending transition of the Flooded Locations and Simulated Hydrographs (FLASH) project from research to operational application (Gourley et al., 2016). FLASH is a suite of real-time tools that use rainfall observations to force hydrologic models to predict flash floods. Two examples of the types of forecast guidance products included in the FLASH project are shown in Figure 1. Potential users include forecasters at both the national and regional scales in the United States, including, but not limited to, National Weather Service Weather Forecast Offices

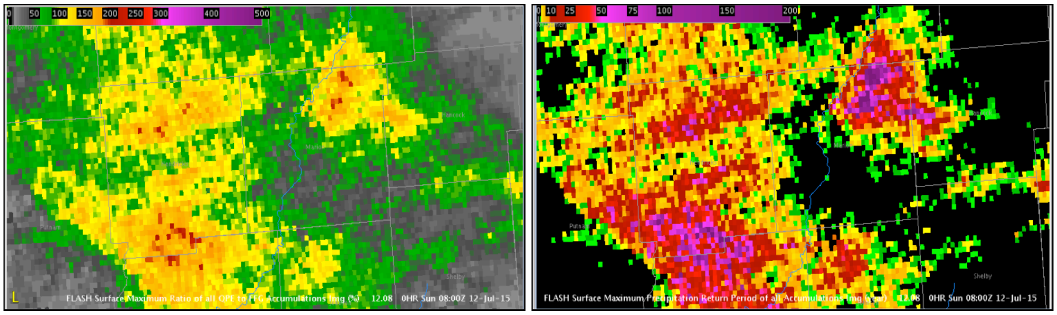
0
50
100
150
200
250
300
400
500
Hancock
Marion
Putnam
Shelby
FLASH Surface Maximum Ratio of all QPE to FFG Accumulations Img (%) 12.08 0HR Sun 08:00Z 12-Jul-15
0
10
25
50
75
100
150
200
Putnam
Shelby
FLASH Surface Maximum Precipitation Return Period of all Accumulations Img (year) 12.08 0HR Sun 08:00Z 12-Jul-15

Figure 1. Two members of the FLASH product suite, the QPE-to-FFG Ratio (Quantitative Precipitation Estimate to Flash Flood Guidance) dynamic visualization (on left) and the QPE Return Period dynamic visualization (on right)

(WFOs), River Forecast Centers (RFCs), and national centers. Both at the national and regional scale, FLASH is designed to assist national forecasters in identifying areas of dynamic flood risk. National forecasters would then work with local forecasters to predict specific threats. When fully transitioned to operations, professional NWS forecasters at offices across the United States will be able to access the decision support tools and use them for situation assessment and judgment justification.

Situation Awareness and Decision Making

Situation awareness is considered to be a prerequisite for decision making, but as of yet, the human factors community has not agreed upon a single, unifying definition. Smith and Hancock (1995) defined SA as “adaptive, externally directed consciousness,” developed through intentional, analytical behavior at an individual level. Likewise, Sarter and Woods (1991) framed SA as the “accessibility of a comprehensive and coherent situation representation which is continuously being updated.” Alternatively, Endsley (1995c) referred to SA as “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future.” The current work draws on Endsley’s (1995c) definition and model due to its widespread acceptance within the weather forecasting operational domain (Jones et al., 2003; Quoetone et al., 2001).

While Endsley’s 1995 Model of SA has received a large degree of attention within the literature, several competing models have attempted to address its perceived limitations. The 1995 Model focused on individual cognition, but as many work environments involve interaction among actors, the Team SA framework was developed to describe information transfers and performance in such situations (Endsley

& Jones, 2001; Salas, Prince, Baker, & Shrestha, 1995). Sensemaking theories, such as the Data/Frame Theory, share some similarities with the 1995 Model of SA (Endsley, 2015b), but they also provide insight into the manner in which decision makers assign meaning and draw conclusions from information (Klein, 2015b). Conversely, the Joint Cognitive Systems (JCS) perspective has envisioned SA as an emergent property within complex systems (Stanton et al., 2006). Although the current work examines SA at the level of the individual decision maker, weather forecasting occurs within a sociotechnical system, and as such, alternative perspectives on SA may provide additional insight.

In addition to the weather forecasting domain, SA has been studied in contexts ranging from aviation and air traffic control (Dao et al., 2009; Moore & Gugerty, 2010; van de Merwe, Oprins, et al., 2012), medicine (Levin et al., 2012), driving (Endsley & Kiris, 1995; Ma & Kaber, 2005; Moore, 2009), and nuclear power management (Burns et al., 2008). SA has received traction in many operational communities, and is regarded as a means to assess and improve task performance (Jones, 2015). Various assessment techniques frame SA as a measurable construct, and they include, but are not limited to probe-based accuracy and response time measures (Endsley, 1995a; Loft, Morrell, & Huf, 2013), self-report measures (Taylor, 1990), physiological measures (Catherwood et al., 2014; Moore & Gugerty, 2010), and qualitative assessments (Hoffman & Coffey, 2004; Klein, 2015a).

Problem Statement

Endsley and Hoffman (2002) state that maintaining SA is one of the most important components of decision making in the weather forecasting domain. Without an accurate situational model, forecast accuracy and timeliness can suffer, leading to possible negative societal impacts (Quoetone et al., 2001). Endsley's (1995c) model provides a theoretical foundation for understanding SA, and has been widely applied as an explanatory device in the weather forecasting domain (Bowden & Heinselman, 2016; Endsley & Hoffman, 2002; Quoetone et al., 2001; Trafton & Hoffman, 2007). What is less understood, however, is how situational uncertainty affects SA.

Previous research has indicated that a gap exists in the knowledge related to the relationship between SA, decision support system design, and weather forecasting. Minotra and Burns (2015) recommended further study of SA within uncertain and dynamic sociotechnical systems, and we propose that the weather forecasting environment is an ideal example of this. Given that uncertainty proliferates within the weather forecasting system, the question then arises of how to accommodate decision makers in ways that promote accurate SA and decision selection. Findings from multiple domains suggest that decision support systems may promote the development of accurate SA, thereby improving operators' abilities to make informed decisions. In weather prediction tasks, forecasters operate on what is sometimes termed "the forecasting funnel," meaning the time-uncertainty continuum (in plain language, the further away in time a forecaster is from a weather event, the more uncertainty there is inherent in what will actually happen). This tension between managing uncertainty, building situation awareness, and the impact that decision support technology has on the two, is the topic of this research.

Significance

In the United States' weather and climate prediction system, accurate and timely weather prediction requires effective interactions among a number of stakeholders. Forecasters are often responsible to emergency management personnel, broadcast media partners, and members of the general public. A loss of SA in the forecasting stage may translate into negative effects as information is transmitted to decision makers at various levels. Indeed, between the years of 1934 and 1999, flash floods occurred at least once per year across the United States causing property and crop damages with an increasing trend (Pielke, Downton, & Barnard Miller, 2002). Flash floods also threaten human life, with several recent events including the 2013 Boulder, Colorado flash flooding (National Weather Service, 2014b) and 2013 Oklahoma City, Oklahoma flash flooding (National Weather Service, 2014c). By examining SA in tasks involving decision making under uncertainty, we will be able to explore the role of uncertainty in the situation assessment process. In doing so, we will also be able to develop guidelines for the user-centered design of forecasting decision support systems.

User-centered design of forecast decision support systems may improve forecast accuracy and lead time (Bowden, Heinselman, Kingfield, & Thomas, 2015) and reduce forecaster workload (Karstens et al., 2015). In a survey of professional forecasters, media personnel, and emergency managers, Morss, Demuth, Bostrom, Lazo, and Lazrus (2015) found that situation awareness (or lack of it) could be transferred among decision makers through risk communications. In the context of the development of the FLASH system, this work contributes to an understanding of behavioral aspects of the flash flood forecasting process. Applying this new knowledge to practice may increase the likelihood that forecasters will be able to use the guidance products effectively during situation assessment to build their own SA, develop their long-term mental models, and execute timely decisions.

Research Questions

This dissertation focuses on resolving four interrelated research questions with the shared goal of contributing knowledge related to SA, decision support technology, and flash flood forecasting. Three studies, discussed in the following chapters, employed quantitative and qualitative research methods in order to address these questions. Each research question addresses a unique aspect of situation awareness in weather forecasting, and are presented below:

- How does data aggregation in a FLASH visualization affect user performance in terms of signal detection, task completion time, and congruence in decisions for a flash flood prediction task? (RQ1)
- How do forecasters build and maintain situation awareness while working under the constraints imposed by uncertainty leading up to a flash flooding event? (RQ2)
- Which tools did forecasters use, in combination and individually, to build situation awareness? How did their SA requirements change at different points along the forecasting compound warning decision process and at different environmental activity levels? (RQ3)
- How is SA influenced by recommender automation at different processing levels during a weather forecasting task? (RQ4.1) To what degree are eye tracking measures (total fixation duration, mean fixation time percentage, time to first

fixation, and mean number of fixations) able to predict situation awareness? (RQ4.2)

Hypotheses

The first research question examined the effects of data aggregation algorithms on signal detection within one of the FLASH guidance products. For the particular decision support visualization, the original design employed a data aggregation technique in order to present a large dataset on a human-interpretable map. With an eye towards understanding how data aggregation affected Level 1 SA (perception) and Level 2 SA (comprehension), we questioned how choice of data aggregation technique would affect performance in terms of signal detection, response time, and likelihood of correctly identifying a threat within the visualization. Based on previous research related to focal attention and visualization design (Pirolli and Card, 1999; Hoffman, Detweiler, Conway, and Lipton, 1993), we hypothesized that the type of data aggregation technique would affect signal detection with the particular FLASH visualization.

The second and third research questions sought to investigate behavioral patterns among forecasters during situation assessment. A focus group methodology was used to explore the relationship between situation assessment and uncertainty management (RQ2); as this was an exploratory, qualitative study, we did not express any testable hypotheses. Conversely, a time- and frequency-based analysis of forecaster behavior related to forecast guidance usage addressed the third research question. Here, we hypothesized that information-seeking behavior during situation assessment would differ across forecast timeframes (watch phase versus warning phase) and across environmental activity levels.

Whereas the first three research questions primarily addressed Level 1 and Level 2 SA, the final research question investigated the effects of decision support automation across all three levels of SA. In other domains, high levels of automation have often been associated with low levels of SA (Kaber & Endsley, 1997). Likewise, we hypothesized that the forecasting decision support automation would lead to lower levels of SA. Additionally, this work assessed the ability of eye tracking measures to predict an individual's amount of SA. Several studies have suggested that eye tracking can accurately predict SA in air traffic control tasks (Moore & Gugerty, 2010; van de Merwe, van Dijk, & Zon, 2012), but to our knowledge, the current work is the first attempt to validate eye tracking as a predictive measure in the field of weather forecasting. Based on previous studies, we hypothesized that eye tracking measures would predict SA.

Scope

The current work is bounded by several delimitations. While many decision support systems are used throughout the weather and climate domain, this work limits itself to the human-centered design and evaluation of the Flood Locations and Simulated Hydrographs (FLASH) suite of guidance products. Although end users could come from a variety of populations, the FLASH products are primarily intended for use by NWS forecasters; as such, this research is focused on human behavior at the level of the individual forecaster. However, we posit that findings would be generalizable to decision makers in environments which involve integration of information sources.

Finally, each study concentrates on situation awareness specifically in flash flooding situations. To some degree, hydrologic forecasting requires some different forms of expertise than other types of weather forecasting, but general behaviors are understood to be similar. Developing systems that support SA development would not only benefit forecasters issuing flash flood watch and warning products, but findings could also be implemented in systems that present information related to other weather threats.

Summary

Situation awareness is a critical component in dynamic decision making processes (Endsley, 1995c). In the weather domain, loss of SA among forecasters can contribute to increased workload as well as reduced lead time and spatial accuracy in emergencies (Quoetone et al., 2001). However, from a theoretical perspective, SA in weather forecasting is not fully understood; indeed, previous research has identified a gap in the knowledge related to SA development under uncertainty (Minotra & Burns, 2015). Previous studies have found that technology can provide support for situation assessment and mental model building (Andra, Quoetone, & Bunting, 2002; Endsley & Kiris, 1995; Kaber & Endsley, 1997; Trafton & Hoffman, 2007). The current work explores the interactions between decision support tools and SA in the weather forecasting domain. As such, we aim to contribute both practical recommendations for weather forecasting decision support systems as well as a theoretical account of the effects of decision support on SA in uncertain, dynamic decision making tasks.

This dissertation begins with a literature review over situation awareness theory, SA assessment methods, and weather forecast decision making. The chapters following the literature review describe three interrelated studies and are written as standalone research topics to be submitted as individual scholarly publications. Using qualitative and quantitative research methods, the studies examine interactions among weather forecasting, situation awareness, and decision making under uncertainty. In Chapter 3, we investigate the relationship between data aggregation in a static flash flood prediction visualization and signal detection. Based on the results, we present evidence-based recommendations for future visualizations using data aggregation. At the time of this work, the material presented in Chapter 3 was in submission as a standalone journal article. Chapter 4 contains a discussion regarding a mixed methods analysis of forecasters information-seeking behaviors during the watch and warning decision making process. This analysis not only resulted in new insight related to situation assessment under uncertainty, but it also revealed information about specific information requirements for building SA in flash flood forecasting. Following this section, Chapter 5 presents results from an experiment that assessed the effects of a type of decision support automation on forecaster SA levels in a flash flood forecasting task. In addition, in Chapter 5, we discuss the adequacy of eye tracking as a predictive measure of SA in weather prediction tasks. This work concludes with a general discussion in Chapter 6, which ties current findings to existing literature.

Chapter 2: Literature Review

On a daily basis, weather forecasters apply meteorological expertise and analytical ability to evaluate threats to life and property. Current understanding suggests that forecasters extract and integrate information from a variety of decision aids in order to build situation awareness and reach a decision about the environmental risks (Trafton et al., 2000). However, forecasters regularly face challenges related to interpreting and using complex data (Doswell, 2004; Pagano et al., 2014), using automated decision support effectively (Karstens et al., 2015; Pagano et al., 2014), and maintaining situation awareness (Hoffman & Coffey, 2004; Trafton & Hoffman, 2007).

Situation awareness (SA) has many definitions, but it is widely regarded as a prerequisite to successful decision making (Adams, Tenney, & Pew, 1995; Durso & Gronlund, 1999; Endsley, 1995c, 2015b; Hoffman, 2015; Wickens, 2015). As an operational concept, situation awareness (SA) has utility for communicating a critical aspect of the weather forecasters' decision processes (Jones, 2015). From a theoretical perspective, although research has considered the development and evaluation of situation awareness (SA) models for more than twenty years, the role of imperfect information and uncertainty in the situation assessment process remains largely unexplored. Overcoming this limited knowledge will be of utmost importance in order to provide forecasters with decision support systems that promote human-system integration as well as SA development. This chapter explores existing literature related to situation awareness, weather forecasting, and decision support technology in complex sociotechnical systems.

The following discussion begins by defining SA as framed by several models of situation awareness and sensemaking. In addition, the discussion will address research methods commonly used to assess SA in operational and experimental contexts. The discussion will also review findings regarding their methodological strengths and weaknesses. After establishing the state-of-the-research in terms of SA, components of the weather forecasting sociotechnical system will be examined with an emphasis on the role of the human forecaster. Finally, outcomes from human factors, meteorological, and decision making research will be synthesized in order to inform user-centered designs for future weather forecasting decision support systems.

Theoretical Models of Situation Awareness and Decision Making

Perhaps due to its longstanding presence in decision making research, SA has a number of definitions within several explanatory models. SA has been framed as a process and product of dynamic cognition in relation to individuals (Chiappe, Strybel, & Vu, 2012; Endsley, 1995c; Smith & Hancock, 1995), groups of individuals (Chiappe, Rorie, Morgan, & Vu, 2012; Salas, Prince, Baker, & Shrestha, 1995), and sociotechnical systems (Stanton et al., 2006). SA has also been closely linked to aspects of attention, memory, and judgment, including sensation, cue detection, monitoring, and comprehension (Hoffman, 2015). For an excellent review of SA theories, refer to Salmon et al. (2008). With several differences in theoretical underpinnings, various models of SA provide unique perspectives into decision making and human performance in complex systems.

Situation Awareness Theories

Sarter and Woods (1991) define awareness in terms of "the accessibility of a comprehensive and coherent situation representation which is continuously being updated in accordance with the results of recurrent situation assessments." This early definition of awareness sets the stage for later theories of SA, such as Endsley's (1995c) Model of SA. Smith and Hancock (1995) took a slightly different perspective, defining SA as "adaptive, externally directed consciousness," which emphasized the view that SA and behavior are driven by ecological factors. Furthermore, Smith and Hancock (1995) caution against models that frame SA as a component of working memory or a type of mental model, arguing in favor of a well-structured and empirically supported definition of SA.

Several explanations of SA center around Neisser's (1976) perceptual-action cycle. The perceptual-action cycle frames cognition and perception as a process in which objects, information, schema, and human behavior are interrelated. Citing the perceptual cycle model, Adams et al. (1995) hypothesize that SA is also a function of perception, memory, and human performance. Similarly, Smith and Hancock (1995) discussed SA as a construct that aligns with Neisser's (1976) perceptual-action cycle, and that SA is equally as important to decision making as attention and workload.

Endsley's 1995 Model of SA. Perhaps one of the most widely cited models of SA, the Endsley 1995 model describes SA at an individual-level, introducing three components of cognition that relate to SA in decision making (Endsley, 1995c, 2000, 2015b; Salmon et al., 2008). Endsley (1988a, 1988b, 1995a, 1995c, 1997, 2000) defines SA as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." Although the Endsley 1995 model has received criticism for failing to distinguish between SA as a product and SA as a process (Baxter & Bass, 1998; Chiappe, Strybel, et al., 2012; Salmon et al., 2008; Stanton et al., 2006), Endsley's (1995c) early work emphasized that the model described both in tandem. In situation awareness—the product component—SA is described as a measure of knowledge, and is furthermore "only that portion pertaining to the state of a dynamic environment." Thus, an individual may hold additional knowledge within memory, but if it is irrelevant to the task at hand, it does not count as SA. The process component, situation assessment, incorporates the processes involved in acquiring and updating SA (Endsley, 1995c). Additionally, SA is viewed as distinct from decision making processes and action choice processes; each occur at different points along the decision making timeline and are governed by different cognitive structures (Endsley, 1995c, 2015b; Wickens, 2015). For the purposes of the present discussion, the term "SA" will be used to refer to both product and process, and when meaningful, will indicate if process or product is more relevant.

In the Endsley 1995 model, presented in Figure 2, the decision making process is cyclical and dynamic, with SA undergoing updates as an environment changes over time (Endsley, 1995c, 2015b). Endsley has proposed that SA is comprised of three distinct levels: perception, comprehension, and projection. In order to acquire and maintain SA, a decision maker must perceive individual environmental elements, which he or she must then compile in order to make sense of the whole picture. If the decision maker is able to perceive and comprehend the meaning of the current state of the

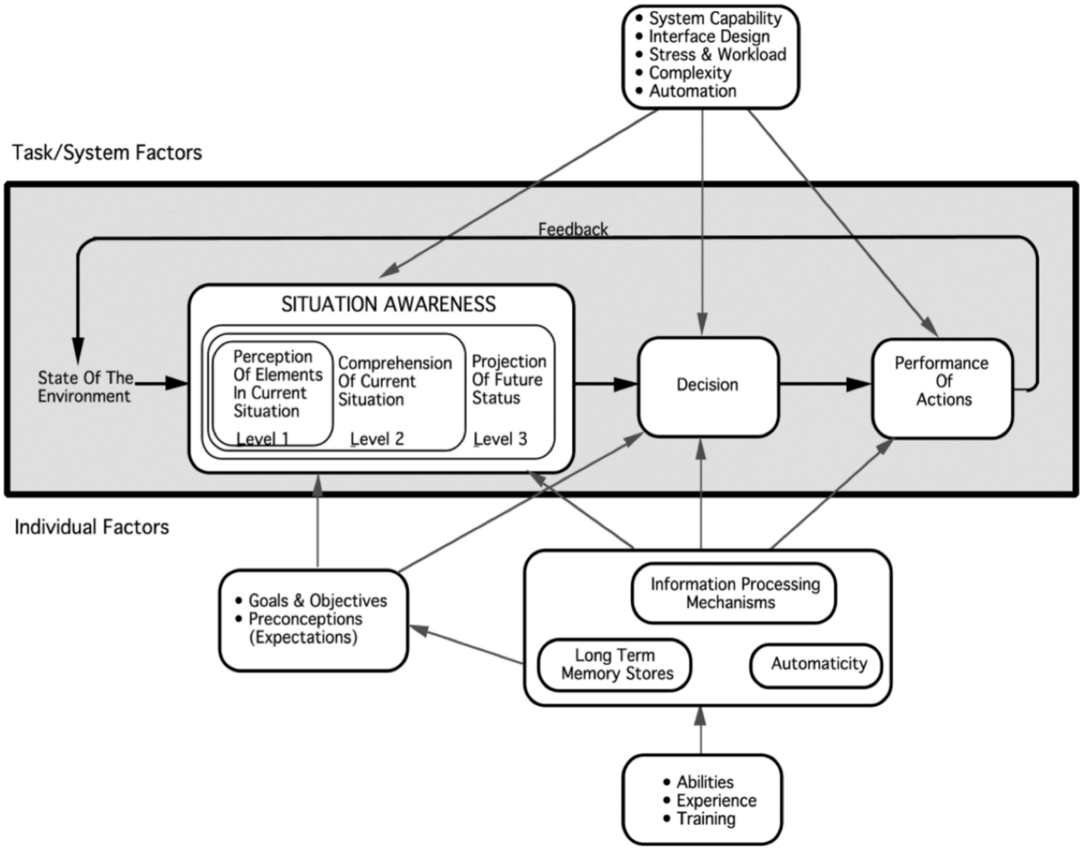

Figure 2. The 1995 Model of SA, proposed by Endsley (1995c)

system, he or she must then be able to envision the future system state based on the current state. When navigated successfully, the decision maker should have a high level of SA.

SA is but one of many factors that drive human decision making. As such, it is still possible for a decision maker to choose poorly or execute a decision incorrectly. While SA influences decision making, SA itself is influenced by individual factors including the operator's goals, expectations, training, and experience; system factors such as system design, workload, task complexity, automation; and individual cognitive factors such as long term memory, attention, and additional information processing structures (Adams et al., 1995; Endsley, 1995c). An operator's ability to acquire SA can be limited by working memory capacity, attentional capacity, and presence of suitable mental models (Endsley, 1995c, 1997).

Despite its comprehensive approach to SA in decision-making, the Endsley 1995 model has received criticism on numerous fronts. In a recent set of articles, Endsley (2015a, 2015b) attempted to rectify several of the most common critiques. Perhaps most prominent is the argument that Endsley's model fails to account for the environment in which an individual operates, and that the model takes a Cartesian, "in-the-head" view of SA (Chiappe, Strybel, et al., 2012; Chiappe, Strybel, & Vu, 2015; Salmon et al., 2008; Stanton, Salmon, Walker, & Jenkins, 2009; Stanton et al., 2006). In a more extreme position, van Winsen and Dekker (2015) question whether it is meaningful and even possible to study SA at an individual level, and instead support a joint cognitive systems approach as an alternative to individual and team SA. Dekker, Hummerdal, and Smith (2010) argue that the Endsley 1995 model studies cognition independently from the environment, neglecting ecology in which the individual operates—essentially, focusing on the awareness, regardless of the situation.

The "in-the-head" criticism can be traced to the debate between triadic and dyadic perspectives within cognitive science (Flach, 2015). With roots in information processing studies, dyadic perspectives seek to understand how internal processes relate to external outcomes (such as how memory affects decision making). Conversely, triadic perspectives frame research questions in terms of the relationship between internal processes, agent characteristics, and ecological properties. For a more complete discussion on the history of this philosophical debate, refer to Flach (2015). Endsley

(2015b) refutes that the 1995 model takes a Cartesian, dyadic perspective, pointing to the inclusion of task and individual factors in the decision making model.

For all its criticism, the Endsley 1995 model was one of the first to frame SA in the context of attention and its relation to perceptual learning of dynamic information (Hoffman, 2015). The model has been generalized to a diverse set of domains, and although it is sometimes construed as a dyadic, information processing based model, it is precisely this generalizability that lends itself to use as an explanatory framework for both dyadic and triadic perspectives (Flach, 2015).

Team SA. Many work-related tasks and environments involve interaction between multiple individuals. In light of this, interest in group decision-making and SA has grown considerably over the years. As with individual-centered models of SA, a collection of theories of team SA exist. Endsley (1995c, 2015b) defines team SA as "the degree to which every team member possesses the SA needed for his or her job." In this explanation, team members operate as individuals while in coordination with each other. The unit of analysis is still the individual, and team SA in effect represents the degree of overlap between each team member in terms of SA. Team SA can exist in dynamic environments wherein individuals' goals may adapt to the needs of the team, thus leading to changes in the extent of overlapping SA requirements (Salas et al., 1995). Alternatively, Dekker (2000) defines crew situation awareness as "the extent of convergence between multiple crew members' continuously evolving assessments of the state and future direction of a process." As with individual human error, some explanations of team SA processes have attributed team errors to the failure of a team as a cohesive entity to maintain SA within individual members (Endsley, 1995c; Kaber &

Endsley, 1998). A breakdown in team SA occurs when even a single team member lacks the amount of SA required to fulfill their role within the team.

Models of team SA that focus on interactions between individuals tend to represent the concept by integrating models of individual SA, team characteristics, and teamwork processes (Endsley, 1995c; Endsley & Jones, 2001; Salas, Fiore, & Letsky, 2013; Salas et al., 1995). With the Endsley 1995 model as a foundation, Salas et al. (1995) proposed that team SA was a function of individual SA and communication within the team. Shown in Figure 3, Salas's (1995) model points out the links between individual characteristics and information processing mechanisms and teamwork processes. In addition, Salas et al. (1995) emphasized the importance of understanding team SA in the context of overlapping knowledge as opposed to the study of multiple individuals' SA levels.

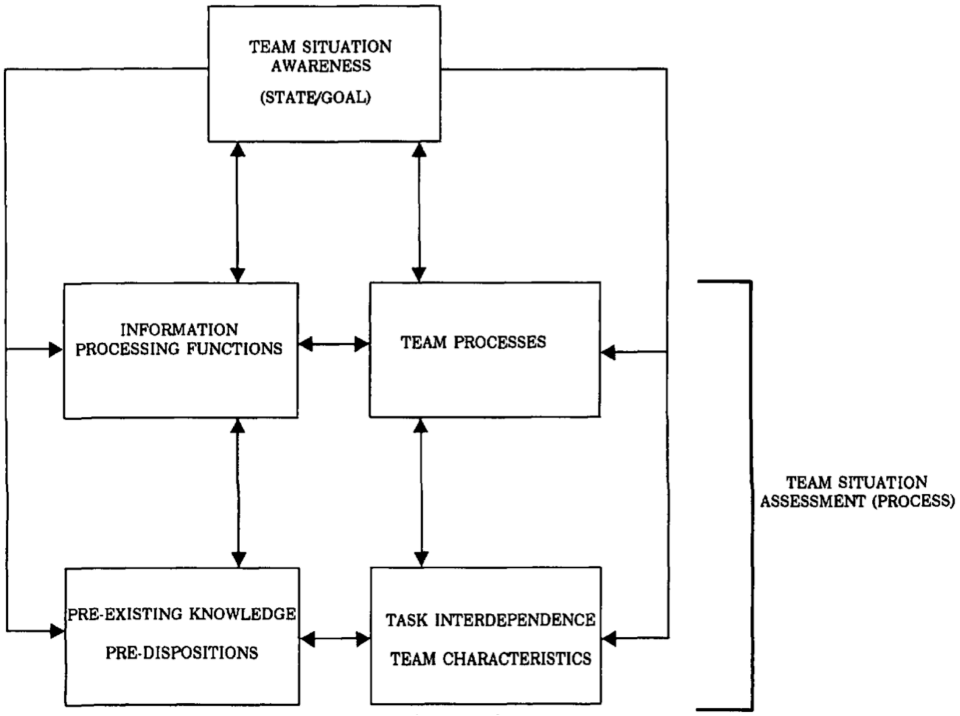

Figure 3. Model of team SA, as shown in Salas et al. (1995)

While definitions of team SA are often tied to the Endsley 1995 model or other individual-focused definitions of SA, additional research has examined mechanisms and processes that facilitate multi-actor decision making. In addition to individual-level situation assessment processes, Salas et al. (1995) suggested that team SA is developed over time through processes involving information gathering, leadership, and fluid communication between team members. Endsley and Jones (2001) expanded upon this framework, introducing a model that explains the team situation assessment process in terms of SA requirements, information processing mechanisms at the team level, communication devices, and workflow processes. Discussions of team SA requirements have aligned with the three levels contained in Endsley's 1995 Model of SA. When translated into teamwork, SA involves tasks related to perception of relevant information, comprehension of one's own goals as well as those of other team members, and the projection of team members' behaviors (Endsley, 2015b). While many researchers have emphasized the importance of communication to the development of team SA, the relationship between team SA and other measures of teamwork, including team attitudes and behaviors have received less focus (Salmon et al., 2008).

An extension of team SA is the concept of shared SA, which refers to the degree that information is shared based on goals shared between team members (Dekker, 2000; Endsley, 2015b; Endsley & Jones, 2013; Salas et al., 2013; Salas et al., 1995). Although scholars occasionally interchange the terms, the role of goals distinguishes the concepts of team SA from shared SA; in team SA, SA is measured as the level of SA possessed by individuals with differing SA requirements within the team, but in shared

SA, it is conceptualized as the overall amount of SA shared by individuals with similar SA requirements within the team (Kaber & Endsley, 1998; Salas et al., 2013). Endsley and Jones (2001) define shared SA as "the degree to which team members have the same SA on shared SA requirements." In an ideal scenario, team members with shared goals each possess the same required knowledge to support their tasks, facilitating a coordinated effort. When team members with similar SA requirements have unequal levels of awareness, teams can become uncoordinated, leading to performance decrements (Kaber & Endsley, 1998). However, shared SA has come under fire in the literature for lacking clarity and for its irrelevance to teams within complex systems (Salmon, Stanton, Walker, Jenkins, & Rafferty, 2010; Stanton et al., 2006).

Some criticism of the individual-level models of team SA originates from literature on distributed cognition approaches to SA (Chiappe, Rorie, et al., 2012; Salmon et al., 2008; Salmon et al., 2010; Stanton et al., 2006). Salmon et al. (2010) claim that current definitions of shared SA are murky at best, and question whether it refers to team members each possessing exact replicas of SA, or simply possessing relevant portions of the total situational picture. The debate between individual-level and systems-level SA scholars may require further discussions in order to arrive at a common ground. Discussions predating the systems-level SA approaches establish that operators can share SA even when individual SA is not identical. Endsley and Jones (2001) state that "the mental models of two team members do not need to be identical, as each member has different functions, nor is it likely they will be" and that effective teamwork can occur as long as "they have enough commonalty to allow comprehension and projection regarding actions that affect each other's tasks."

In further evidence of Endsley (2015b), Bolstad, Riley, Jones, and Endsley (2002) discuss a case in which two military officers had different end goals, and each comprehended the same Level 1 SA elements differently. Yet, despite the differences between the two operators, they shared Level 3 SA requirements; both used their different views of the same situation to project a shared situational outcome (Bolstad et al., 2002; Endsley, 2015b).

Mutual SA, a potential remedy to this debate, has received some support in the literature, and is a concept that describes when multiple operators are aware of each other's SA (Chiappe, Rorie, et al., 2012; Shu & Furuta, 2005); put simply, each knows what they individually know and they both know that the other knows, too.

Nevertheless, there are several aspects of team SA that remain yet undiscovered. In particular, how team SA changes over long periods of time is not well understood, especially when teams are engaged in projection activities (Salas et al., 2013). Furthermore, while several studies have determined that shared displays and shared mental models improve team SA and team performance (Endsley, 2015b; Endsley & Jones, 2001), some scholars have called for the need to address issues related to how perception and comprehension affect projection of the future (Wickens, 2015).

Distributed Situation Awareness. Grounded in the field of distributed cognition (Artman, 2000), schema theory (Bartlett, 1932), genotype/phenotype schema, and Neisser's (1976) perceptual cycle model of cognition, the theory of Distributed Situation Awareness (DSA) describes team SA from a systems-level perspective (Salmon et al., 2008; Salmon et al., 2010; Stanton, Salmon, Walker, & Jenkins, 2008; Stanton et al., 2006). DSA originated from the supposition that teamwork involved different cognitive processes than individual properties. Thus, an alternative but complementary model was needed to describe SA in complex sociotechnical systems (Stanton et al., 2006). Prior SA theories represented team SA as a state of knowledge within the heads of individual team members, but Stanton et al. (2006) developed a model of SA based on propositional networks of distributed knowledge.

DSA describes awareness as an emergent property from complex systems in which human and non-human actors interact (Salmon et al., 2008; Stanton et al., 2006). In DSA, ownership of SA moves away from the limited domain of individual operators and into the broader realm of the system in which many operators exist.

In order to understand the differences between team SA and DSA, it is important to acknowledge several underlying assumptions of DSA. From its inception, DSA was intended as an alternative approach to team SA, which was seen as too restrictive in its focus on individual cognitive mechanisms (Salmon et al., 2010; Stanton et al., 2006). Stanton et al. (2006) developed DSA as a means to explain the development of SA in dynamic, sociotechnical systems in which humans and technology both possess forms of SA. The construct of SA in distributed systems was redefined as "activated knowledge for a specific task within a system" (Salmon et al., 2008; Stanton et al., 2006). However, this is not to negate the utility of individual SA models. Individual SA theories of SA development, such as the Endsley 1995 model, may in fact occur during an individual actor's performance; however, the DSA concept seeks to understand SA in terms of linkages between actors and the non-human elements within the system.

DSA centers upon a set of theoretical propositions; namely, that SA can be held both by human and non-human system members, that several agents may have different interpretations of the same situation, and that communication activities are key to acquiring SA within the system (Stanton et al., 2006). As shown in Figure 4, Salmon et al. (2008) represent the emergent SA with the largest circle, showing information transactions between human and nonhuman actors with arrows. As depicted in their model, while individuals may indeed possess SA, they argue that the most meaningful SA occurs when knowledge is activated in dynamic complex networks. In line with

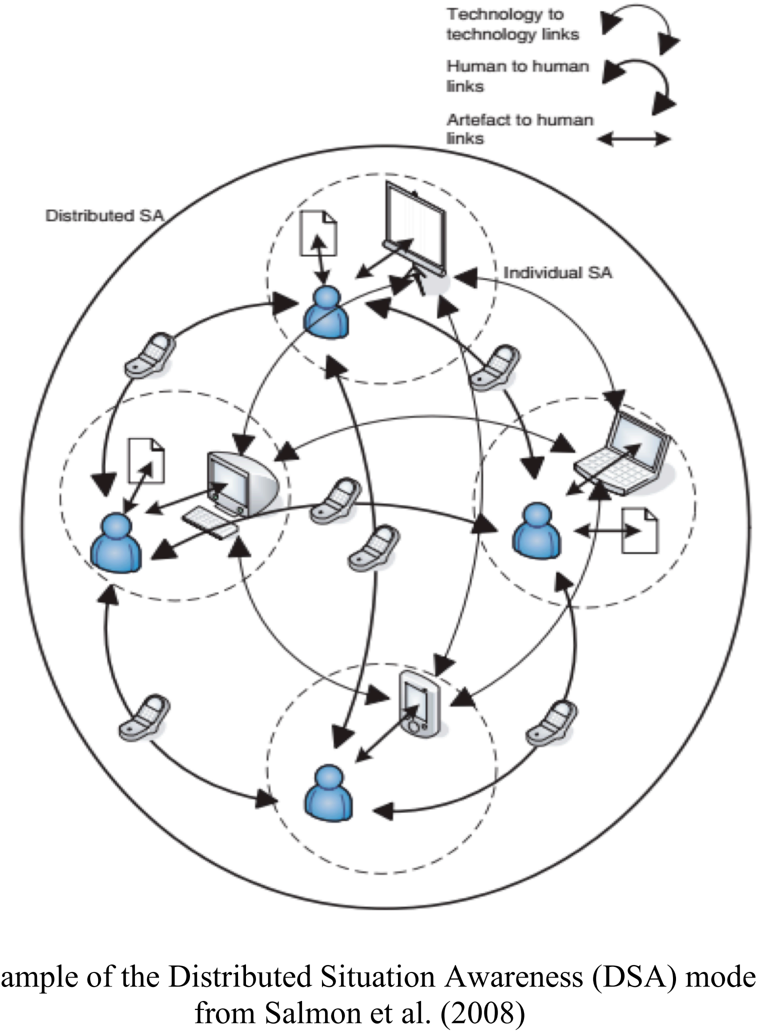

Figure 4. Example of the Distributed Situation Awareness (DSA) model, adapted from Salmon et al. (2008)

concepts such as shared SA, DSA supports the notion that system agents may have overlapping SA when their goals have similar SA requirements, and that system actors with varying levels of individual SA may counterbalance the DSA of the entire system (Stanton et al., 2006).

However, according to the DSA perspective, shared SA and DSA do not refer to the same concept. According to Stanton et al. (2006), although they may measure the same SA at times, shared SA refers to instances in which system agents share both goals and SA requirements, and in DSA, agents have different but compatible purposes and SA requirements. Salmon et al. (2010) questioned the meaning of shared SA as proposed by Kaber and Endsley (1998). Specifically, they refer to uncertainty regarding whether the concept refers to individual agents "sharing" in the sense that each agent has identical SA, or in the sense that each agent holds a unique but relevant piece of the whole picture (Salmon et al., 2010). Instead of a shared approach, Salmon et al. (2010) suggest using *transactive SA* and *compatible SA* constructs for understanding a system's DSA. Whereas transactive SA refers to the process in which system actors exchange information, compatible SA refers to the phenomenon in which system agents hold distinct components of system awareness based on differing information sources, yet are compatible due to overlapping operator goals. The existence of compatible SA is what unites members of distributed systems (Salmon et al., 2010; Stanton et al., 2009; Stanton et al., 2006).

DSA has received several criticisms from proponents of individual-level SA. The concept of SA as an emergent property has been met with skepticism. In a review of the DSA model, Endsley (2015b) voices the concern that the DSA perspective overlooks valid and time-honored definitions of what SA is, and confuses SA with the nonhuman entities that may be used to gain it. However, this critique is countered by the supposition that in complex systems, SA represents activated knowledge and is an emergent property that cannot be reduced to the individual level (Salmon et al., 2010; Stanton et al., 2006). While DSA supporters argue that SA is primarily meaningful when knowledge transactions occur, they also suggest that human operators have drastically different backgrounds, and thus cannot create the same situation model even after encountering the same information (Salmon et al., 2010). However, this view has met with some discomfort, even from supporters of distributed cognition approaches to SA. Chiappe, Rorie, et al. (2012) state that this view opposes current understandings of human perspective-taking and ability to share one's intentions. Such debates show that though distributed approaches to SA are fairly recent developments, a sharp divide between the individual-level and systems-level perspectives has already appeared. Both concepts may have utility, but in future research, it will be necessary to distinguish which applications and goals are more appropriately modeled at each level.

Situated Situation Awareness. The Situated approach to situation awareness, proposed by Chiappe, Strybel, et al. (2012), is a relative newcomer to the SA debate. As with DSA, Situated SA is based on a distributed cognition perspective, but instead identifies the individual as the appropriate unit of analysis (Chiappe, Strybel, et al., 2012). Citing memory constraints and criticisms of the Endsley 1995 model, particularly the perceived Cartesian "in-the-head" approach and the product versus process debate, Chiappe, Strybel, et al. (2012) justify an alternative explanation of individual SA. The situated approach utilizes the three levels of SA included in the

Endsley 1995 model, but introduces the idea that the most efficient processing strategy involves using the environment as its own representation, instead of storing all relevant information in internal models as suggested by Endsley (Chiappe, Strybel, et al., 2012; Chiappe et al., 2015). In the situated approach, SA is not maintained entirely by internal cognitive processes but by off-loading SA into props within the environment; props could include checklists, mnemonic devices, automated reminders, or any number of memory aids. Thus, the boundaries defining SA expand from an individual alone into the inclusion of ecological entities that an individual encounters while acquiring SA. Further work extends the framework of individual distributed cognition into team settings (Chiappe, Rorie, et al., 2012).

Early studies of SA have suggested that working memory capacity limits the amount of relevant information that can be used in situation assessment, but that long-term memory structures aid SA through mechanisms such as mental models and schemata (Adams et al., 1995; Endsley, 1995c). Chiappe, Strybel, et al. (2012) believe that phenomena like change blindness and perception failures indicate that focus on internal representations is overemphasized in current understandings of SA. Chiappe, Strybel, et al. (2012) view SA as a synthesis of two explanatory models of sensemaking: the Construction-Integration Theory (CI) of sensemaking (Durso, Rawson, & Girotto, 2007; Kintsch, 1988) and Relevance Theory (Wilson & Sperber, 2002). Based on premises from these theories, Chiappe, Strybel, et al. (2012) posit that factors, including ease of encoding, frequency of information use, ease of access, and individual factors such as expertise and working memory capacity affect whether SA is stored as an internal or external representation.

Proponents of the Situated SA model argue that SA is distributed, or off-loaded, into structures in the environment in order to reduce load upon working memory; however, this premise has received sharp criticism from others in the field. Endsley (2015b) rejects the notion that long-term memory plays a minimal role in SA, emphasizing that a distributed cognition approach is not appropriate for an individual processing situation. In the words of Endsley (2015b): "Information that exists in the environment… but of which the operator is not aware… does not constitute SA. It is by definition information of which he or she is not aware."

Although partially disputing this assessment, Chiappe, Strybel, et al. (2012) acknowledge that off-loading may not always be the most effective strategy, stating that, "individuals must incorporate external representations into their operations in a way that increases the likelihood of successful performance." Chiappe et al. (2015) emphasize activated knowledge as being key to understanding the situated approach. Indeed, knowledge that an operator is not aware of may not be SA, but situated SA is instead created when the right information is activated at the right time. In other words, knowledge may be present within props, but such knowledge does not translate into SA until it is activated by the operator (Chiappe et al., 2015).

The central premise to the situated approach—that SA cannot be contained entirely within working memory—is actually expressed as a condition in the Endsley 1995 model of SA (Endsley, 1995c, 2015a, 2015b). Endsley (2015a) offers the commentary that the situated approach is based on a set of studies that use novice operators as participants, yet several studies have shown that experts are able to use mental models and schema more effectively than novices. Despite framing itself as an alternative to “in-the-head” views of SA, the situated approach is not as opposed to more traditional models of SA as how the concept has been theorized (Endsley, 2015a). Acknowledging that working memory and relevance limit storage capacity, Endsley (2015a) questions how best to identify the quantity and content of information, as well as the moments in time in which knowledge activation must occur for representations to be used for awareness.

Sensemaking

While the concept of sensemaking has many similarities to situation awareness, it offers a unique perspective on decision making processes. Sensemaking has been present in the literature since at least the mid-twentieth century, but Weick (1995) brought renewed attention to it within organizational contexts (Klein, Phillips, Rall, & Peluso, 2007). At individual and team levels of decision making, sensemaking models have been applied to domains including fire ground command (Klein, 1993; Klein, Calderwood, & Clinton-Cirocco, 1986), military command and control (Jensen, 2009; Klein, 1989), weather forecasting (Pliske, Crandall, & Klein, 2004), air traffic control (Malakis & Kontogiannis, 2013), and intelligence analysis (Pirolli & Card, 2005). As a field of study, sensemaking is often associated with naturalistic decision making methods (Klein, 2008, 2015a).

Klein et al. (2007) define sensemaking as “the deliberate effort to understand events,” and is often associated with an initial condition of surprise; that is, a subject will engage in sensemaking purposefully when a situation does not match his or her expectations (Klein et al., 2007; Weick, 1995). Klein, Moon, and Hoffman (2006a) report that sensemaking integrates several cognitive processes, including curiosity, comprehension, mental model creating, and SA. In addition, several models represent sensemaking as a function of problem detection and identification (Klein, Moon, & Hoffman, 2006b).

Although it contains elements of many recognized aspects of cognition, sensemaking has the greatest intersection with mental model construction (Klein et al., 2006a). Klein and colleagues (Klein, 1999, 2008; Klein et al., 2006a, 2006b; Klein et al., 2007) represent the concept as an iterative process in which decision makers attempt to comprehend stimuli and events in order to identify an appropriate action to take. That this perspective has similarities to several definitions of situation awareness has not escaped scholars. Klein et al. (2006a) recognized the commonalities between sensemaking and SA, but stated, "in contrast [to Endsley's SA product], sensemaking is about the process of achieving these kinds of outcomes, the strategies, and the barriers encountered."

While this perspective falls prey to the SA product versus process debate, sensemaking may then be likened to situation assessment. Nevertheless, there are several distinctions that make it a unique and valuable concept for understanding decision making. Sensemaking is often represented as a retrospectively-driven process in which a decision maker makes sense of the present based on past events (Weick, Sutcliffe, & Obstfeld, 2005). Although sensemaking involves retrospective analysis in large part, it also has a forward-looking component, similar to Level 3 SA, in which the ultimate goal is to determine an appropriate action in the context of the situation (Weick et al., 2005). However, unlike comprehensive models of SA, sensemaking theories tend to be constrained to the activities involved in problem detection and comprehension.

The following discussion will present two explanations of sensemaking. While other theories exist, this section focuses on those theories put forth by scholars who have connected the concepts of sensemaking and decision making to situation awareness.

Recognition-Primed Decision Model. The Recognition-Primed Decision (RPD) model, proposed by Klein et al. (1986), provided groundwork for several theories of decision making, including the Endsley 1995 model of SA (Endsley, 1995c). Klein (1993) recognized that in some decision scenarios, decision makers experience significant limitations in terms of time and resources. Analytical decision making, in which a decision maker evaluates several alternatives, often requires enough time to compare and contrast the options; in time-constrained situations, this may be a luxury that one does not have. The RPD model takes an adaptive approach, asserting that experience and iterative evaluation play a role in finding a workable solution. Similar to the decision feedback loop contained in the Endsley 1995 model of SA, the RPD model explains aspects of situation assessment; however, it excludes cognitive processes involved in comparison of alternative choices (Klein, 1993). This recognitional model addresses rapidly made, expertise-driven decisions; additionally, it provides a framework for understanding what has been known as intuitive decision making (Klein, 1989, 2015a).

According to the RPD model, shown in Figure 5, sensemaking in such situations involves two processes, situation assessment and action assessment (Klein, 1999). When a decision maker is initially subjected to a situation, they may find it familiar and typical, or in the case when sensemaking is needed, they may find something atypical

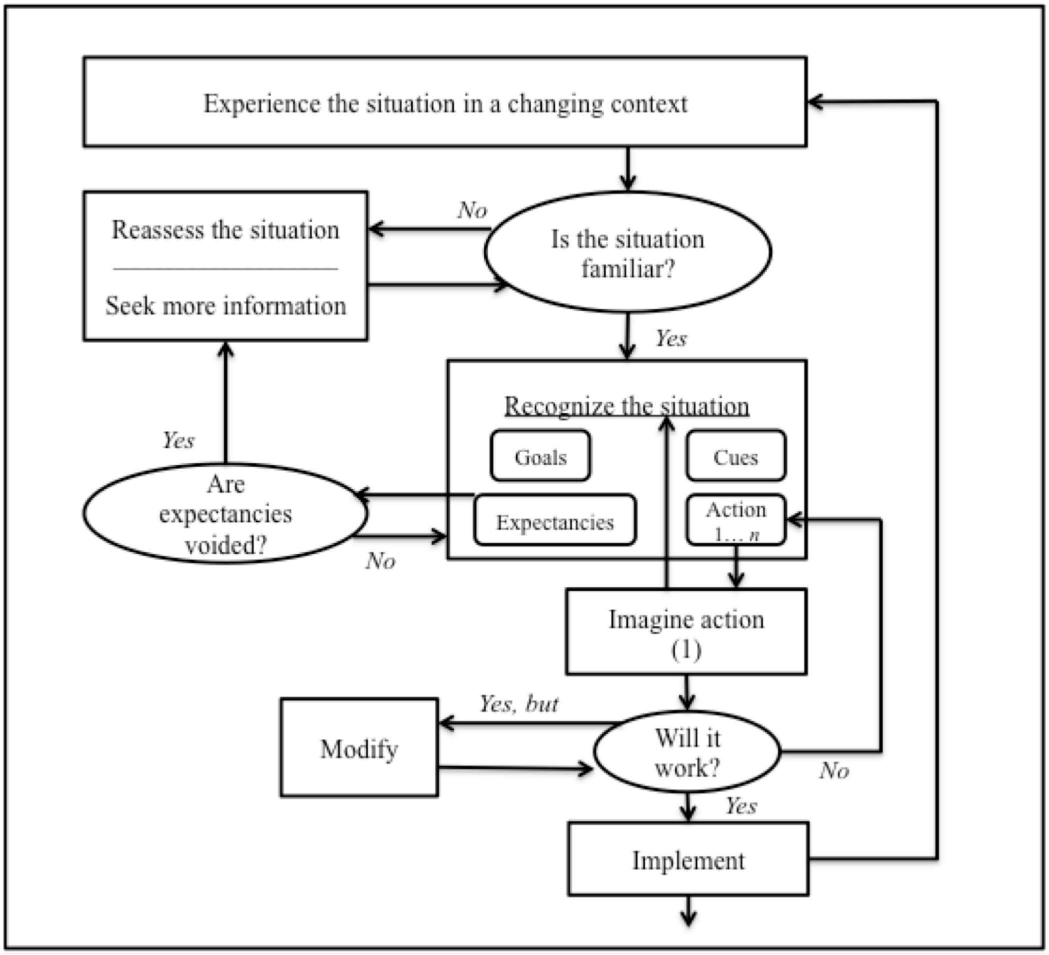

Figure 5. Recognition-Primed Decision (RPD) model

and surprising (Klein, 1993). In order to identify solutions to an unfamiliar or unexpected situation quickly, decision makers imagine potential actions based upon their goals, expectancies, situational cues, and past experience. Then, potential actions are subjected in order of occurrence to rapid analysis. Cognitive processes like mental simulations internalize and thus speed up the decision making process. Once the first workable action—the satisficing solution—is found, the decision maker can implement it. Thus, the solution either resolves the situation, or the decision maker receives feedback and can reassess the situation. Pattern recognition and experience plays a critical role in recognition-primed decision making (Klein, 1989).

The focus on satisficing solutions is one of several characteristics that distinguish the RPD model from models of SA. Instead of choosing the best decision, Klein (1989) suggests that in some scenarios, it may be more efficient for a decision maker to choose the first functional solution. Klein (1993) found that when decision makers lacked relevant experience, they rarely had the mental models required for conducting mental simulations of potential solutions. Indeed, novice decision makers lacked the mental models required to generate action choices rapidly or accurately, a finding which was later supported by Endsley (1995c). This recognitional approach provides a framework for understanding decision making by experts under great time pressure and great uncertainty (Klein, 1993).

Data/Frame Theory of Sensemaking. After establishing the RPD model, Klein and colleagues (Klein et al., 2006b; Klein et al., 2007) sought to understand deliberate sensemaking processes, encapsulating their findings in the Data/Frame Theory of sensemaking. Like the RPD model, the Data/Frame Theory presents a focused view of intentional and conscious sensemaking; it goes beyond previous frameworks in its attempt to explain how people construct and interpret data. Existing models of sensemaking and SA, including the Endsley 1995 model, placed great importance on the role of data; however, as models focused on data processing, the data itself had rarely been studied closely. Klein et al. (2007) argued that prior efforts to explain sensemaking and situation awareness had neglected to define how data is identified, and set forth the Data/Frame Theory to explain processes involved in data construction and interpretation. The Data/Frame Theory presents sensemaking as a "closed-loop sequence between mental model formulation (backwards) and mental stimulation (forwards)" (Klein et al., 2006b).

Central to the Data/Frame Theory are the concepts of *data* and *frames*. According to Klein et al. (2007), a *frame* is "an explanatory structure that defines entities by describing their relationship to other entities" and which provides a "structure for accounting for the data and guiding the search for more data." Frames can be likened to Neisser's (1976) schema concept, a cognitive construct involving attention, memory, and experience in order to direct information management. While frames can take a variety of forms, mental models are perhaps the closest to being the primary variety (Klein et al., 2006b). Klein et al. (2007) distinguish between two types of frames that play different roles in sensemaking: "just-in-time" frames and "comprehensive" frames. With "just-in-time" frames, interpretation of data is based on basic, assumed knowledge of the data elements. "Comprehensive" frames are those in which data are interpreted based on knowledge of complete relationships between data elements. In weather forecasting, a "just-in-time" frame could be likened to a member of the public looking at a radar image and recognizing that the representation indicated severe weather. In the same situation, a "comprehensive" frame would be one held by a professional forecaster, whose mental models of the weather would contain knowledge about atmospheric and environmental relationships that could promote further severe weather.

Compared to the environmental elements perceived in Level 1 SA (Endsley, 1995c), Klein et al. (2007) proposed that "data" is a relative concept. In the Data/Frame Theory, data are abstractions of elements in the environment. In this definition, in order to be understood, one must not only consider the actual state of the environment, but also perceptive and abstractive cognitive processes that shape what becomes data (Klein et al., 2007). This alternative perspective, based on doubt in the information-processing models, supports the idea that stimuli and events are rarely perceived without introducing individual bias.

The Data/Frame Theory and the Endsley 1995 model of SA address related but distinct concepts. Whereas Endsley's (1995c) framework uses cognitive structures to explain how individuals perceive and make use of information, Klein et al. (2007) commits to a model in which frames are used to synthesize data and draw meaning from them. While both explanations contain a forward-looking component, there are several critical differences involved in their structures and applications.

Sensemaking involves two concurrent processes in which frames define what data are, and conversely, data determine the construction and selection of frames (Klein, 2015b). Shown in Figure 6, the Data/Frame Theory posits that sensemaking consists of a series of cyclical processes. Initially, when a decision maker is exposed to a situation in which they must make sense of some information, her existing frames (such as mental models) will determine which informational elements are relevant to the situation at hand—these become the initial data set. Concurrently, the perceived data will also be used to determine which frame is most appropriate. Identifying a frame is followed by a series of synthesizing processes, in which the decision maker elaborates, preserves, questions, and reframes data. In some situations, data may be incomplete, leading to gaps in the decision maker's knowledge. In the elaboration cycle, the

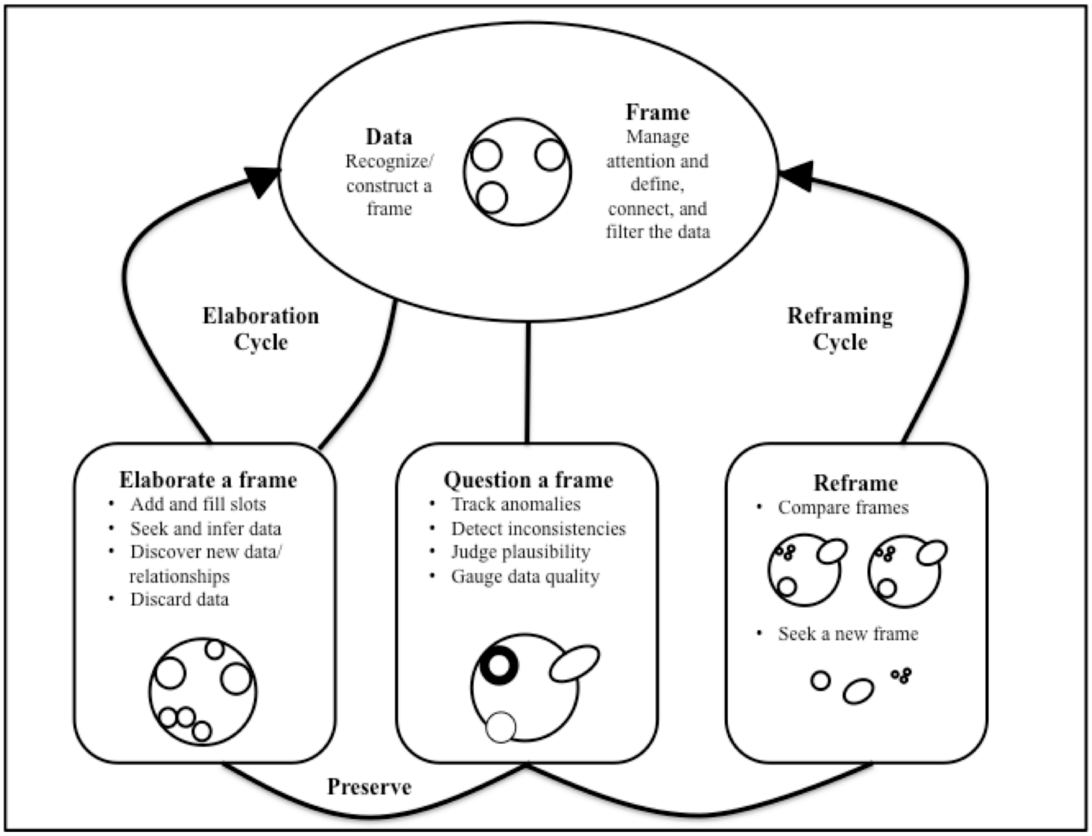

Figure 6. Data/Frame Theory of Sensemaking, adapted from Klein et al. (2006)

decision maker seeks out additional data or removes irrelevant data in line with the frame(s) in use. The decision maker may also question the frame in use if anomalies in the data are found, or if data is of poor quality; this may be due to an imperfect frame choice. If, in fact, the frame choice was adequate, but for some reason the decision maker senses inconsistencies between it and the data, she may seek to preserve the frame by engaging in further elaboration. This activity may uncover additional relevant data that can then be used to update the frame in use. Conversely, it may be more appropriate to reframe the data completely if the original frame is a poor fit for the data after a close analysis (Klein et al., 2006b; Klein et al., 2007).

As with the RPD model and the Endsley 1995 model of SA, the Data/Frame Theory posits that expertise plays a large role in the sensemaking process. Klein et al. (2007) argue that expert and novice decision makers use essentially identical procedures when engaged in sensemaking, but disparities in performance are due to differences in expertise. After many experiences working through the Data/Frame cycles, experts build up extensive collections of frames, whereas novices have relatively few frames in their repertoire. Over time, novices develop their frames, adding to their quantity and quality, and in addition, reframing when necessary. The Data/Frame Theory has received criticism for failing to explain how cognitive processes and structures integrate to form, develop, and use frames. Frames have proven to be a contentious idea, and even the proposing authors acknowledge limitations in knowledge about this construct (Chiappe, Strybel, et al., 2012; Klein et al., 2006b). The usefulness of the Data/Frame concept has also come under fire in the SA literature. Proponents of the Situated SA approach argue that the Data/Frame Theory has little utility for explaining SA due to its failure to separate long-term knowledge from short-term, situation-centered knowledge (Chiappe, Strybel, et al., 2012). In addition, Endsley (2015b) has argued that it is inappropriate to focus so deeply on recognitional approaches to sensemaking, and has gone so far as to suggest that the Data/Frame Theory has few explanatory advantages over the Endsley 1995 model of SA. While this opinion may overlook several benefits of the Data/Frame Theory, it is true that the framework neglects to identify the processes involved in recognizing when a situation merits analytical or recognitional decision making approaches.

In response to Endsley's (2015b) critique, Klein (2015b) responds that the Data/Frame Theory was never intended to be a comprehensive model. Its focus is on deliberate sensemaking in uncertain situations, and it represents a closer look at data construction than what the Endsley 1995 model provides (Klein, 2015b). Where Endsley (2015b) states that the Endsley 1995 model and Data/Frame model each address problem detection, Klein (2015b) argues that the Data/Frame model adds value by incorporating reframing processes. Where Endsley (2015b) argues that the Data/Frame model and Endsley 1995 model each describe data gathering and interpretation, Klein (2015b) responds that the Endsley 1995 model deals with data gathering and synthesis, but not construction, which is the purview of the Data/Frame model. Clearly, both models offer insight into the decision making process, despite contention produced in scholarly debate.

Summary of the Models

The aforementioned models of SA and sensemaking provide a framework for understanding SA and decision making across multiple levels of analysis. Sensemaking theories and SA models complement each other in several ways. Sensemaking explicitly addresses understanding of situations in which uncertainty exists; indeed, one would rarely need to engage in sensemaking if there wasn't uncertainty. Few SA models address how uncertainty management fits into the decision making process, or how its existence affects situation assessment. In this way, sensemaking offers much to understanding the concept of SA.

Not all scholars fully accept the significance of SA as a factor in decision making and human performance. Dekker and Hollnagel (2004) refer to current explanations of SA as “folk models,” in which phenomena are essentially not measurable and therefore explained through substitution and overgeneralization. Instead, they call for an increased focus on explaining human decision making in terms of performance, as in their view, the joint cognitive system is much more meaningful than the study of human cognition separated from its ecological situation (Dekker & Hollnagel, 2004; van Winsen & Dekker, 2015).

Several questions emerge from a synthesis of the SA and sensemaking literature. Crosscutting the different perspectives on units of analysis for SA is the representation of SA as functional understanding of a situation that results in action choice and performance (Chiappe, Rorie, et al., 2012; Endsley, 1995c; Klein, 1989). There may be utility in viewing SA at multiple levels of analysis, and further discussions would be necessary to establish appropriate frameworks for discussing SA in individual cognition as well as across sociotechnical systems. At the individual level of analysis, knowledge related to SA and human performance is lacking; for example, although a number of assessment techniques for SA exist, it is still unclear how to distinguish between inaccurate and incomplete SA (Baxter & Bass, 1998). Likewise, it is important to question how an individual’s SA accuracy affects decision making and performance, which should shed light on the relationship between SA and situational uncertainty (Minotra & Burns, 2015).

Assessment of Situation Awareness

Over time, a variety of methods and tools for assessing situation awareness have emerged from the research community. The following section presents an overview of some of the most commonly used approaches in the literature.

Probe-based Techniques

Probe-based techniques are perhaps one of the most commonly used methods to assess SA. Several probe-based methods have been discussed in the literature, each one in turn addressing a different facet of SA. Probes, designed based on expert knowledge of the human subjects' workflow, assess SA in terms of absolute accuracy; either subjects demonstrate SA, or they do not. Probe-based techniques have been criticized for assuming a situational ground truth, against arguments that this may not be true for all scenarios.

Endsley (1988b) proposed the Situation Awareness Global Assessment Technique (SAGAT). SAGAT has been validated in a variety of domains, including airfield combat (Endsley, 1988b, 1995a; Endsley, Selcon, Hardiman, & Croft, 1998a), air traffic control (Jones & Endsley, 2004), emergency medicine (Levin et al., 2012), and driving performance (Ma & Kaber, 2005). SAGAT is a technique in which a subject is asked questions (also known as probes) while undergoing a simulated scenario; this is sometimes referred to as an on-line method. At randomized intervals throughout the simulation, the scenario is paused and all relevant displays are temporarily cleared; at this point, the probes are presented. Once all the probes are answered, the scenario is started from the pause point and runs until the next set of probes are due to begin. When the entire simulation is complete, a composite score of performance is calculated from a comparison of responses to the ground truth in the scenario. SAGAT responses can also be categorized into sub-scores associated with components of SA (Endsley, 1988b, 1995a).

Similarly, the Situation Present Assessment Method (SPAM) is an alternative technique that adopts a real-time approach to assessing SA (Chiappe et al., 2015; Durso et al., 1995). Like SAGAT, SPAM presents queries to participants in a simulated scenario and evaluates response accuracy against a ground truth. SPAM is distinguished from SAGAT, though, in its application of the probes and the inclusion of response time as a valid predictor of SA (Dao et al., 2009; Durso et al., 1995; Loft, Morrell, & Huf, 2013). Instead of freezing the simulation, the SPAM procedure presents probes without pausing the workflow. In principle, this allows subjects to access information as and when it is queried, demonstrating SA when the subject is aware where information is stored in the environment. SPAM's probe technique makes it an accessible assessment tool for proponents of the situated approach to SA. By allowing subjects to access the information components, it inherently assumes that SA is not just what can be contained within the head, but that which can be stored using environmental and task-related cues (Chiappe, Strybel, et al., 2012).

Comparisons of SAGAT and SPAM have produced evidence both for and against real-time and freeze-time probing techniques. In a study of chess players, Durso et al. (1995) evaluated the efficacy of SAGAT and SPAM in predicting players' SA levels; they found that while both methods were viable measures for SA, response accuracy was a significant predictor for SAGAT, but not for SPAM. Conversely, response time was a significant predictor for SPAM, but less so for SAGAT (Durso et al., 1995). However, in a similar comparison of real-time and SAGAT probes, Jones and Endsley (2004) found a correlation between accuracy-based probes and response time-based probes, albeit a weak one. Furthermore, Jones and Endsley (2004) also identified a weak correlation between workload measurements and the real-time probes, which indicates that further work is needed in order to identify performance effects from probe-based techniques.

In line with this finding, some practitioners have expressed concern that SAGAT is that memory limitations may diminish SAGAT's ability to measure SA. It has been argued that by freezing the scenario, probes measure recall instead of overall SA (Dao et al., 2009). However, referencing a study by Endsley (1994), Durso and Gronlund (1999) state that no significant difference was found in SAGAT scores measured at 20 seconds and 6 minutes after freezing the scenario. This finding suggests that concerns regarding memory limitations may be overstated, but it also assumes that memory does not decay significantly after 20 seconds (Durso & Gronlund, 1999).

Dao et al. (2009) attempted to overcome limitations of SAGAT and SPAM by combining aspects of both techniques into one method. In the combined probe method, participants were given access to the displays while answering the probes, but after the simulation had ended, so as to not affect mental workload (Dao et al., 2009). This approach, along with other real-time probe-based techniques, provides a view of at least a portion of SA (Adams et al., 1995). However, this view is partial at best; probes only reflect the components of SA that they directly query, and thus, reflect SA in the form of performance, and only indirectly represent the underlying cognitive processes. In addition, conclusions from on-line probes are difficult to generalize past the scenario for which they were designed, and further problems can occur when simulated scenarios are not realistic (Adams et al., 1995). This can be overcome by designing scenarios around real situations, such as historic events from the real world (Adams et al., 1995).

Rating Approaches

Similar to probe-based techniques, ratings-based approaches for assessing SA can be designed in more than one form. Most often, SA is assessed through subjective rating methods, such as the Situation Awareness Rating Technique (SART), or through observer rating methods.

The SART is an easy-to-use method that presents a questionnaire of subjective items to subjects at the end of a simulation in order to assess their own level of SA (Taylor (1990) as cited in Selcon, Taylor, and Koritsas (1991)). SART consists of a series of questions that assess ten components of a subject's SA; the questionnaire can be administered either at the completion of the simulation, or intermittently during pauses in the simulation, as done with SAGAT (Endsley, Selcon, Hardiman, & Croft, 1998b; Selcon et al., 1991). While completing the questionnaire, subjects rate each item on a seven-point scale (Selcon et al., 1991). Components of SA considered in the SART are related to perceptions related to cognitive demand, availability of attentional resources, and the subject's understanding of the present situation. While requiring much less subject matter expertise than probe-based methods, some studies have shown that the SART is sensitive to background experience and task difficulty (Selcon et al., 1991). From the ten components included in the SART questionnaire, a total score can be calculated which is designed to reflect a subject's overall SA.

An alternative type of rating measure has subject matter experts observe and rate operators' SA. SA is assessed against a set of behaviors that are associated with high and low levels of SA, often developed from a grounded knowledge of workflow and task demands. In the Situational Awareness Linked Indicators Adapted to Novel Tasks

(SALIENT) method, observers watch operators perform their tasks while recording relevant operator behaviors using the checklist (Muniz, Stout, Bowers, & Salas, 1998). The SALIENT method was developed as a means to assess team interactions and SA, and so the checklist focuses not only on individual actions, but also interpersonal interactions and information handoffs (Muniz et al., 1998). However, as with operator-generated rating techniques like SART, observer ratings have received criticism of their subjective nature. Self-guided ratings have been viewed as unreliable and even inappropriate. Indeed, the issue of how much trust should be placed in a subjective score of SA has been raised, calling into question the degree that an individual can truly be aware of his or her own knowledge (Salmon et al., 2010). An individual may believe that they have a high level of SA, but as Endsley (1995a) points out, many significant safety failures have occurred even when operators believe they are behaving appropriately. Likewise, observer rating methods are subject to the same critiques.

Methodological validation studies consistently show that other assessment techniques perform more reliably than SART (Loft et al., 2013). Problems related to predictive power and timing bias may affect outcomes. Endsley (1995a) cites poor correlation between SART and SA performance measures, having previously suggested that a positive or negative situational outcome could bias a subjective rating if presented at the end of a scenario (Endsley, 1988b). In relation to probe-based techniques, Loft et al. (2013) also found that SPAM exhibited stronger predictive power than SART in relation to performance in a submarine track management task. Finally, Salmon et al. (2009) identified a significant correlation between SAGAT scores and operator performance, but failed to find a correlation between SART and performance.

Observational Approaches

While many quantitative methods for assessing SA exist, alternative perspectives on SA can be produced using qualitative approaches. A deeper understanding of SA can be gained through a variety of observational methods such as cognitive work analysis, naturalistic decision making (NDM) research (Klein, 2008), the critical decision method, and propositional network modeling (Salmon et al., 2010), among others.

In order to address perceived shortcomings of laboratory-based evaluations of decision making, the NDM perspective emerged as an alternative assessment method in the field of SA (Klein, 2008). The NDM framework assesses situation awareness as a component of a decision making process situated within a specific ecology; thus, this approach shifts assessment out of the laboratory and into the environment in which decisions are made (Klein, 2008). The naturalistic approach has been closely linked to the RPD model and theories of sensemaking, providing critical structure for understanding these phenomena in a qualitative manner (Klein, 2008). Often employing interview-based approaches, NDM studies have uncovered a wealth of information undiscoverable through more empirical methods; NDM-based literature has explored decision decision-making in contexts including, but not limited to, weather forecasting (Pliske et al., 2004; Smallman & Hegarty, 2007), fireground command (Klein et al., 1986; Klein, Calderwood, & MacGregor, 1989), and military command (Klein et al., 1989).

The NDM framework has spawned a variety of techniques for eliciting knowledge regarding decision making from people engaged in the environment in question. Semi-structured interviews, such as the critical incident technique, are often used to conduct a post-event analysis with key decision makers (Klein et al., 1989; Randel, Pugh, & Reed, 1996). While some studies analyze decisions following a real-world event, such as a building fire (Klein et al., 1986), others employ a hybrid approach that blends laboratory-based controlled scenarios with post-event interviews (Randel et al., 1996; Smallman & Hegarty, 2007). Studying electronic warfare technicians in the United States Navy, Randel et al. (1996) used the critical incident technique to elicit information about the decision making processes that participants used during the critical incident, defined as an event in which a successful outcome in the simulation is dependent upon the participant's behavior during the event. Interviews assessed decision making and situation awareness through structured queries related to one or more of the critical incidents, an unstructured discussion of the entire simulation, a discussion of the simulation's timeline, and an identification of key decision points and factors affecting the participant's decision (Randel et al., 1996). The critical decision method is a derivation of Flanagan's (1954) critical incident technique, and it focuses the interviews on decisions made during the scenarios as opposed to controlled incidents (Hoffman, Crandall, & Shadbolt, 1998).

Work analysis is a widely used method that facilitates the evaluation and modeling of complex sociotechnical systems. The cognitive work analysis (CWA) approach, discussed extensively by Vicente (1999), allows designers to identify environmental and cognitive constraints that influence work demands on system resources. Not only has CWA been used extensively for system modeling, but it has also lent itself well to understanding SA and decision making in complex systems

(Minotra & Burns, 2015). The procedure often targets multiple aspects of the work environment, using methods such as analyses of the work domain, decision ladders, cognitive strategies, organizational and social transactions, and worker cognitive competencies (McIlroy & Stanton, 2011; Read, Salmon, Lenné, & Stanton, 2014; Vicente, 1999). In combination, outcomes from these analytical components create a map detailing relationships between system entities and resources as well as requirements for successful decision-making and performance.

Observational methods have long been used to shed light on decision-making processes within individuals and teams. Recent applications of qualitative methods include modeling and assessing complex sociotechnical systems. In relation to the DSA model, propositional network modeling uses verbal protocol analysis, hierarchical task analysis, and the critical decision method to develop a representation of information transactions between agents within the system (Salmon et al., 2010; Stanton et al., 2006). However, due to its qualitative nature, propositional network modeling is limited in its ability to assess the quality and quantity of operator and overall system SA; such assessments must be based on subjective measures based on observer judgments (Salmon et al., 2010). The propositional network modeling technique is part of the broader Event Analysis of Systemic Teamwork (EAST), another methodology based on the NDM framework, which has been used to identify SA requirements in distributed sociotechnical systems in many contexts (Stanton, Salmon, & Walker, 2015; Walker et al., 2006).

Although NDM methods have the potential to provide insight into decision-making, several limitations exist to their effectiveness. First, although they elicit rich sets of knowledge related to operator behavior and task demands, models developed through such means are often very context-specific and thus difficult to generalize to broader applications. Second, testing and validating models is also a challenge. Finally, from a data collection standpoint, the relationship between SA and decision processes is still not well understood, and a failure to account for this could affect NDM findings. Indeed, while SA affects decision outcomes, it has also been suggested that SA also affects selection of decision-making strategies (Endsley, 1997; Minotra & Burns, 2015). A deeper understanding into the mechanisms associated with SA and decision-making is needed in order to advance the NDM methods as SA assessment measures.

Physiological Indices

While several probe- and rating-based approaches have been correlated to SA performance, these methods have only been shown to be effective in controlled laboratory environments (Moore & Gugerty, 2010). In order to overcome the limitations associated with direct and indirect measures, physiological measures have been identified as potential predictors of SA. Several physiological measures have been evaluated in relation to their ability to predict SA, including eye movements and electrical brain activity.

As a measurement method, eye movement analysis has had a surprisingly long association with SA, being both lauded and criticized for the information it has the potential to provide. Tracking eye movements may provide a more direct way than probe-based methods for evaluating perceptual processing when developing SA (Adams et al., 1995). Although early attempts to link SA to eye movements failed to identify differences in levels of SA (Durso et al., 1995), recent efforts to gauge the effectiveness of the method have produced more positive results (Moore & Gugerty, 2010; Yu, Wang, Li, & Braithwaite, 2014). In a study of air traffic controllers, Moore and Gugerty (2010) found that the number of eye fixations was a significant predictor of SA in terms of probe response accuracy; the number of fixations was inversely related to the number of errors that a participant made when answering probes.

In a flight simulator-based study of pilots using a head-up display (HUD), Yu et al. (2014) measured the number and duration of eye fixations and compared them against self-reported perceived workload and subjective measures of perception and SA. In their procedure, SA was determined in an observer rating approach. During the simulation, the experimenters would randomly switch on a warning light; if the pilots reacted correctly, the experimenters recorded the participant as having "high SA" and if an incorrect or no response was taken, the experimenters recorded the participant as having "low SA" (Yu et al., 2014). While Yu et al. (2014) found a correlation between mental workload and SA, the appropriateness of a binary measure of SA should be called into question. Yu et al. (2014) state, "Pilots who were able to identify the activated warning light have better SA performance and show significantly lower workload." This association has been supported by other studies, but a binary, observer-based judgment of SA seems to be more an artifact of the experimental design than a true measure of SA.

Electroencephalography (EEG), a measure of electrical brain activity, has also been evaluated as a measurement technique for SA. EEG has been used to measure SA with some success, but concerns have been raised that while EEG may provide insight into brain activity during situation assessment, the technique, still does not provide a measure of information contained in memory, information completeness, or comprehension level; this critique also holds for any measure of SA based on performance (Endsley, 1995a). Nevertheless, recent applications of EEG measures for SA support explanations of top-down processes involved in SA. In a series of two visual-based perception studies, Catherwood et al. (2014) used EEG to evaluate brain activity during loss of SA in situations with high levels of uncertainty. Using a combination of EEG to measure brain activity and a signal detection-based approach to assess loss of SA, brain imagery revealed several high-order areas of the brain associated with SA. Most notably, the orbitofrontal cortex, an area associated with cognition under uncertainty and stimulus-response contingencies, was activated during loss of SA during experimental tasks (Catherwood et al., 2014). These findings suggest that top-down processes such as memory and mental models can be assessed objectively, despite suggestions otherwise (Dekker & Hollnagel, 2004).

Although physiological measures have shown promise in relation to SA assessment, it is important to consider attentional limitations that may not be captured by such measures. For example, eye tracking may not be able to capture loss of SA due to the change blindness, a phenomenon in which eyes fixate upon a stimulus, but the information is not encoded (Chiappe, Strybel, et al., 2012; Endsley, 1995a; Moore & Gugerty, 2010). However, as Durso and Gronlund (1999) suggest that limitations in the coverage of physiological-based measures may be overcome if used in conjunction with additional SA measures.

Summary of the Assessment Methods

Situation awareness can be evaluated with a number of different methods, including performance-based measures, subjective rating approaches, observational methods, and physiological assessments. Performance-based measures show great promise as predictors of SA; however, Durso and Gronlund (1999) caution that one cannot assess SA by only looking at performance, but there is precedent for using it as an implicit measure of SA. Physiological indices and performance-based measures do not always correlate well to SA (Salmon et al., 2010). Furthermore, questions of methodological validity as well as ability to generate repeatable and meaningful outcomes have been raised with respect to probe-based techniques and NDM (Dekker, 2000). Each approach has strengths and weaknesses; a combination of methodological approaches has been advocated as a way to balance these trade-offs (Dekker, 2000). Existing SA assessment methods explain portions of the phenomena; it is possible that a more comprehensive model of SA in relation to its underlying mechanisms and influencing factors may be gained through methodological triangulation. In order to get the broadest picture of this complex construct, future research should work to assess SA from multiple perspectives, balancing information gained through observational research with findings from performance-based measures, physiological measures, and subjective approaches.

Decision Making in Weather Forecasting

Situation awareness (SA) has gained traction in operational environments, in part due to its ability to facilitate communication between disciplines, to translate cognitive theory into design deliverables, and to develop training systems (Byrne, 2015;

Jones, 2015). The weather forecasting community often speaks about decision-making in terms of SA, and studies of forecaster SA and sensemaking have revealed much about forecast decision processes (Bowden, Heinselman, Kingfield, & Thomas, 2015; Hoffman & Coffey, 2004; Klein et al., 2006b). In studying sociotechnical systems such as the weather forecasting domain, greater understanding of human decision making, mental models, and SA can allow system developers to match technology to the needs of the users (Endsley, 2001). In this way, the study of SA can do much to inform the field of weather forecasting and decision support design.

In order to take advantage of an enhanced understanding of SA and its underlying mechanisms, one needs to recognize the complexities in the weather forecasting domain; as a sociotechnical system, the forecasting domain consists of human and technological agents. At an individual level, forecasting interweaves cognition, interpersonal communication, and technology use. At the systems level, forecasters interact with emergency management personnel, broadcast media, and members of the general public, amongst other system actors. Outside of operational forecasting responsibilities, forecasters may also interact with researchers and environmental modelers. The weather domain is truly a system of systems. While it is possible to look at situation awareness at multiple levels within the system, the following discussion presents a view of information requirements for situation awareness at level of the individual forecaster and the interactions involved between human and technical system elements.

Weather Forecasting from a Human Perspective

Although details may vary between regional NWS Weather Forecast Offices (WFOs), the forecasting process generally remains constant across the United States. This also may hold true at the international level; comparing the work behaviors of weather forecasters in the United States and Australia, Kirschenbaum (2004) observed very similar decision making processes in the forecasters, despite each location using different types of decision aids and technology. For a thorough discussion on the daily workings in a WFO, refer to Daipha (2010) who discusses observational fieldwork conducted over the course of nearly two years. Summarizing Daipha (2010), operating in shifts, forecasters work individually and in small groups to maintain situation awareness over environmental states. Forecasters receive information primarily via computer monitors placed on personal workstations. In the words of Doswell (2004), the influx of information sources available through these modern workstations is like "trying to drink from a fire hose" and that excessive amounts of data can lead to information overload.

Several scholars have described the flow of information through the forecast decision making process. Weather forecasting involves a large amount of visual processing and information synthesis; these are necessary to gain awareness and make sense of unfolding environmental patterns (Daipha, 2010). A number of observational studies have described forecaster information-seeking behavior and integrative reasoning during simulated forecasting activities (Barthold et al., 2015; Heideman, Stewart, Moninger, & Reagan-Cirincione, 1993; Karstens et al., 2015; Morss & Ralph, 2007). Morss and Ralph (2007) found that forecasting ability involved synchronization of a variety of information sources, including computational model outputs, real-time environmental observations, individual background knowledge related to geography and weather patterns, end user needs, and feedback from previous forecasts and other forecasters.

A forecaster's ultimate goal is to maintain awareness over an environmental situation in order to predict weather threats in a timely and accurate manner. In addition to timeliness and spatial accuracy, forecasters have cited low forecast bias and consistency between forecast products as desirable aspects of forecasts (Morss & Ralph, 2007). The NWS definition of a "good" forecast is based on verification statistics, including probability of detection and false alarm ratio (Bowden et al., 2015). Under the current paradigm, if a weather event is forecast but not observed, the forecast is categorized as a false alarm; however, current observation methods may not be able to detect every weather event, leading to false negatives in verification. Recent calls for a renewed look at forecast verification methods have attracted attention, particularly in light of improved understanding of the forecast decision making process (Bowden et al., 2015). An alternative view of forecast goodness holds that a forecast is "good" if it closely matches the forecasters' knowledge and experience, the observed environmental state prior to and during the forecast period, and if a forecast end user gains benefit from the knowledge conveyed in the forecast (Murphy, 1993).

Individual Factors. Weather forecasting is an inherently human-centered activity. Individual characteristics play a large role in the processes involved in and outcomes from forecasting. Highly skilled forecasters possess a number of technical abilities and personal characteristics. Forecasters should be adaptable to new technologies, be able to translate knowledge into actionable information, be able to synthesize numerous information sources, have strong interpersonal skills, and possess knowledge of end user requirements (LaDue, 2011; Stuart et al., 2006). Forecasting ability is also affected by background knowledge, including local geographic and climatological knowledge, and prior experiences with the weather phenomenon in question (Morss & Ralph, 2007).

In terms of background experience, professional forecasters typically hold a Bachelor's degree or higher (Daipha, 2010); however, LaDue (2011) found that few forecasters learn their trade in formal educational settings. Using a grounded theory approach, LaDue (2011) hypothesized that instead of through formal instruction, forecasters learn their skills in interactive environments in which other forecasters essentially mentor less experienced forecasters. Interviews revealed that strong social relationships, regular exposure to weather phenomena, and maintaining a professional identity played a large role in development of forecasting expertise.

Expertise is a key factor that affects decision-making, and technology usage has been suggested as a means to distinguish between non-experts and experts. Using the Critical Decision Method, Pliske et al. (2004) interviewed professional forecasters, finding that non-experts often based decisions solely on numerical models and a set of assessment procedures. Conversely, experts took a more adaptive approach and were more able to integrate personal background knowledge with the model predictions, which may indicate more accurate mental models, a better use of forecaster mental models, or perhaps both. Trafton (2004) defined mental models as a dynamic collection of visual and textual information that allows the subject to draw inferences about spatial and qualitative relationships. Mental models may affect the way in which a forecaster understands the environmental situation. Forecasting errors may occur when a mismatch exists between what the forecaster perceives in the environment and what their mental model would lead them to expect to perceive.

Mental model formation may be affected by visual memory and spatial cognition, two factors that have been associated with effective performance in weather forecasting (Pliske et al., 2004; Smallman & Hegarty, 2007; Trickett & Trafton, 2006). Visual memory and spatial cognition may affect pattern recognition ability. Daipha (2010) suggests that a good visual memory may improve forecast timeliness and spatial accuracy, citing a forecaster's comment that displaying several visualizations on the workstation monitor made the information difficult to distinguish and interpret. The importance of spatial cognition is further supported by Smallman and Hegarty (2007), who identified differences between expert and non-expert forecasters in terms of spatial ability. Forecasters created information displays that they then used to create a forecast for a local airfield. Measures of spatial ability, forecasting background, and feedback on information displays were also taken. From these findings, Smallman and Hegarty (2007) identified an inverse relationship between expertise and complexity of the forecaster-generated information displays. The authors posit that this could be due to novice forecasters expecting to need more context to the situation, whereas experts exhibited stronger performance due to more developed mental models.

Finally, while spatial cognition and forecasting experience are instrumental in the forecasting process, it is important to note that forecasters often require specialized expertise. Forecasters have different SA requirements for different types of weather phenomena. While pattern recognition is an important skill for forecasters, environmental events exhibit different behaviors, and forecasters sometimes train to specialize in forecasting events, such as flooding (Daipha, 2010). A forecaster responsible for flash flood forecasting in the northeastern United States will be trained to recognize a different set of environmental behaviors than the patterns a fire weather forecaster in the southwestern United States would evaluate. Thus, in order to discuss decision-making in weather forecasting, it is highly relevant to discuss the context in which the forecaster is situated.

Environmental Information. Forecasters access a substantial collection of computational models for environmental prediction, which not only provide direct estimates of environmental variables, but also indirect information. When evaluating model predictions, forecasters assess model accuracy and bias, which can result in different information sources being preferred across different geographic locations and under certain environmental conditions (Morss & Ralph, 2007). Individual meteorological and environmental phenomena have corresponding numerical prediction models, though some models can be useful for gaining SA in more than one type of weather event. Meteorological ensemble frameworks, a type of model that generates multiple outputs based on permutations of the input variables, time-lagging predictions, or contrasting independent modeling systems, have been the focus of much research in recent years, and have been shown to improve forecaster confidence in operations (Evans, Van Dyke, & Lericos, 2014). Information needs and tool use may also differ depending on forecast timeframe. Morss and Ralph (2007) observed that forecasters reviewed environmental observations and personal experience more often in forecasts of events occurring in fewer than six hours, but used computational models more often when forecasting events twelve hours to one day in advance.

In flash flood forecasting, prediction methods are often based on rainfall estimates and basin scale. With development beginning in the 1970s, one of the first models for flash flood prediction is flash flood guidance (FFG), a tool based on the amount of rainfall needed to produce flash flooding over a specified land area and timespan (Clark, Gourley, Flamig, Hong, & Clark, 2014). For a more detailed discussion of the history and development of FFG, see Clark et al. (2014). Due to its longstanding use within the National Weather Service, forecasters are accustomed to using FFG, which may affect willingness to adopt more modern methods of flash flood prediction.

Recent efforts to leverage modern data collection technology and crowdsourcing techniques have produced several alternative datasets that have shown early success in terms of forecaster use. Gourley et al. (2013) created a database of flash flood measurements and impacts in an attempt to use the dataset to expand knowledge on societal impacts of flash flooding. The database is comprised of three datasets: an archive of U.S. Geological Survey (USGS) stream gauge measurements across the United States, a record of verified NWS Local Storm Reports (LSRs) related to flash flooding events between 2006 and 2011 and their locations, and finally, a set of flash flood reports collected from members of the U.S. public through the Severe Hazards Analysis and Verification Experiment (SHAVE) between 2008 and 2010 (Gourley et al., 2013). In an extension of this work, Barthold et al. (2015) developed an additional dataset for hydrologic event verification; this set merged NWS LSRs, USGS stream gauge measurements across the United States, and reports of flash flooding collected through a citizen science crowdsourcing mobile application. Together, these datasets are some of the earliest attempts to develop verification and feedback methods for hydrometeorological forecasting (Barthold et al., 2015).

Forecasting Decisions and Feedback. After integrating background knowledge, expertise, meteorological and environmental information sources, and historical datasets, forecasters may then be able to predict a future state of the environment; this may be realized in the issuance of a forecast product, such as a watch or a warning. Bowden et al. (2015) proposed that this is a compound warning decision process, which involves a cycle of threat detection, threat identification, and reidentification. Throughout this process, forecasters update threat predictions as the situation changes over time and space. This framework conceptually aligns with Endsley's (1995c) Model of SA; errors may occur in detection, influencing identification, or even if a forecaster correctly detects patterns associated with severe weather, an insufficient mental model could lead to a misidentification of the threat.

Following issuance of a forecast product, local forecast offices might be able to assess whether or not the predicted event actually occurred, which can then be reported in a collection of verification statistics. Verification datasets can help forecasters to manage uncertainty in the decision making process by providing feedback about the adequacy of past forecasts. Morss and Ralph (2007) found that discussions with end users, including emergency managers, provided valuable information to forecasters, in turn helping them to modify future forecasts to suit user needs and improve forecast accuracy. Similarly, in a survey of NWS forecast offices, Novak, Bright, and Brennan

(2008) found that end users frequently request information related to forecast uncertainty and forecaster confidence; as the forecasting field moves towards a more probabilistic paradigm, communication between actors in the weather domain and comprehensive verification datasets may facilitate situation assessment and awareness.

In a study where participants issued wind speed and visibility forecasts, forecast skill improved not only with a higher experience level, but also as feedback on performance increased (Murphy & Daan, 1984). Likewise, Morss and Ralph (2007) observed that end user feedback affected future products issued by forecasters, and information about forecast quality and value was the most influential, which corresponds to the proposition that feedback is a prerequisite for SA (Sarter & Woods, 1991). It is possible that this feedback loop serves to update forecaster dynamic mental models, as suggested by Trafton (2004).

Situation Awareness in Weather Forecasting

Understanding the weather forecasting decision making process is necessary in order to improve information display technology and decision aids, which in turn should improve forecasting outcomes. A robust integration of human agents into the forecasting system should lead to more timely and accurate forecasts as well as lower workloads placed on forecasters themselves. As an aspect of decision making, much of the existing literature that intersects SA and weather has been situated within the domains of air traffic control (Moore & Gugerty, 2010) and pilot awareness (Bustamante, Fallon, Bliss, Bailey, & Anderson, 2005). A deeper study of SA in the context of weather forecasting can have real and meaningful implications for the design of future forecasting systems and technology.

Imagine a scenario in which a forecaster's goal is to monitor information sources from a geographic region in order to predict flash flooding. From the perspective of Endsley's (1995c) Model of SA, in the perception stage (Level 1), the forecaster must recognize elements in the forecast environment that are relevant to his or her goals. In the case of flash flood forecasting, elements could include model forecasts of anomalous conditions, reports of flooding from verifiable sources, developing weather systems in the surrounding area, environmental conditions conducive to flash flooding, or geographic features specific to an area, such as burn scars. Once key elements are perceived, the forecaster may be able comprehend a deeper meaning from the elements in combination (Level 2); the forecaster may recognize that heavy rainfall over a burn scar is a risk factor for flash flooding. The projection component of SA occurs when the forecaster is able to extend the current state of the environment to a potential future state (Level 3). In this example, the perception of elements and the comprehension that the trend is associated with high risk could lead the forecaster to identify a future timeframe for flash flooding to begin. Extending past the situation assessment process and into the decision and performance stage may include the forecaster choosing to issue a flash flood warning to alert local officials and residents of the impending threat.

Much of the weather forecasting within the United States National Weather Service requires interaction between several levels of the weather enterprise, and thus could also be studied with a team SA framework. In a scenario involving a developing severe weather threat, forecasters not only work cooperatively, but they often work in tandem with emergency management and public agencies to maintain SA throughout the system. Despite disagreement between proponents of individual and team versus systems-level (Stanton et al., 2006) and Situated SA (Chiappe, Rorie, et al., 2012; Chiappe, Strybel, et al., 2012) frameworks, analyzing SA at multiple levels within the weather enterprise may generate meaningful information that could improve understanding of SA in weather decision-making. Each of the four SA theories employs a variety of assessment methods and each evaluates SA at different units of analysis.

Few studies have empirically assessed the SA of weather forecasters, but several have addressed the overall forecast decision-making process. The sensemaking perspective has gained traction in the research community, having been used to explain the information comparison, integration, and problem detection activities used in weather prediction (Klein, Pliske, Crandall, & Woods, 2005; Pliske et al., 2004). Using Comparative Cognitive Task Analysis (C2TA), Kirschenbaum (2004) found that professional forecasters regularly engage in activities related to extraction of information, comparison of information sources, and comparison of the perceived environment to mental models.

Interestingly, studies of weather forecasters have revealed behaviors that throw the widespread acceptance of previous assumptions of expert sensemaking into question. Sensemaking theories often accept that information seekers are swayed by a confirmation bias, but observational research has found that professional forecasters often try to disprove their initial assumptions (Hoffman, Trafton, and Roebber (2006), cited in Klein et al. (2006b)). Forecasters have been observed seeking information in a goal-directed manner, which may be one means of obtaining actionable information from large quantities of complex datasets and displays (Trafton et al., 2000).

In a field work study to assess how forecasters build mental models, Hoffman and Coffey (2004) found that forecasters use a recognition-primed decision-making strategy. Given the importance of pattern recognition in forecaster training, this is a logical finding. In the resulting model of forecaster sensemaking, situation assessment is affected by mental model strength as well as pattern recognition, and in turn, it affects the way in which the forecaster interprets the data at hand (Hoffman & Coffey, 2004).

Design for SA in Weather Forecasting Decision Support Systems

Human factors research has produced a number of design guidelines to improve user performance and human-systems integration. Furthermore, assessment of SA often provides system developers with insight into the design of work systems to match users' decision-making processes (Jones, 2015). Critics of highly automated systems have warned that without integrating knowledge of cognition into the designs of forecasting technology, the human component of forecasting will be lost to the detriment of society (Murphy, 1993). Within the last decade, studies in graph comprehension and information visualization have produced new knowledge that can be applied to the user-centered design of forecasting decision-aiding technology (Hegarty, Smallman, & Stull, 2012; Trafton & Hoffman, 2007; Trickett & Trafton, 2006).

With respect to designing for SA in sociotechnical systems, Endsley (2001) has provided the Situation Awareness-Oriented Design (SAOD) cycle. This three-pronged approach involves an initial evaluation of SA requirements, followed by an iterative process of system design and evaluation. The SA requirements analysis, often conducted as a cognitive task analysis, serves to identify key operator goals and the information needed to accomplish them. In the next stage, SA-oriented design, design guidelines and SA requirement information are used to develop systems from a user-centered perspective. SA-oriented design is a process in which design guidelines are centered on supporting user cognition, SA, and goal accomplishment. After an initial design has been developed, evaluation occurs in the third stage. Any of the assessment methods discussed in the previous section can be used to evaluate the adequacy of the second-stage design; however, Endsley (2001) recommends the use of SAGAT as it provides a quantitative estimate of SA that conceptually links SA to decision choice.

Design Challenges. The complexities of many sociotechnical systems demand a unique approach in order to accommodate the diverse goals and requirements of system actors. Scholars like Endsley (2001), Hoffman and Coffey (2004), and Trafton and Hoffman (2007) have advocated addressing complex system design in terms of challenges as opposed to design-by-rule. Challenges purposefully avoid reliance on rules. For example, a design guideline might state that meteorological decision aids designers should consider that “pastels might work well in certain applications for both backgrounds and target symbols” (Hoffman, Detweiler, Conway, & Lipton, 1993). When designing a system with many users and many goals, a challenge might instead be phrased as, “Support for parallel processing, such as multi-modal displays should be provided in data rich environments” (Endsley, 2001). This transition towards a systems-perspective promotes technology that is adaptable to users; adaptability is critical in systems where users may have conflicting goals and need to use the same decision aids for a number of purposes.

In this way, design for SA in complex systems becomes less about successful task completion and more about adequate information transfer and integration (which should theoretically lead to successful task completion). Principles discussed in the literature that are relevant to decision aid design are numerous, and include the following:

- The Sacagawea Principle: "Human-centered computational tools need to support active organization of information, active search for information, active exploration of information, reflection on the meaning of information, and evaluation and choice among action sequence." (Endsley & Hoffman, 2002)
- The Lewis and Clark Principle: "The human user of the guidance needs to be shown the guidance in a way that is organized in terms of their major goals. Information needed for each particular way should be shown in a meaningful form, and show allow the human to directly comprehend the major decisions associated with each goal." (Endsley & Hoffman, 2002)
- "Direct presentation of higher level SA needs (comprehension and projection) is recommended, rather than supplying only low level data that operators must integrate and interpret manually." (Endsley, 2001)
- "Support for global SA is critical, providing an overview of the situation across the operator's goals at all times… and enabling efficient and timely goal switching and projection." (Endsley, 2001)

In practice, principles such as these may manifest themselves in different ways in different sociotechnical systems. In the weather forecasting domain, consideration of the Sacagawea Principle would promote the development of goal-centric displays and interfaces that facilitate easy exploration of the data. Similarly, the Lewis and Clark Principle encourages design through focusing on users' cognitive processes. Challenges support user-centered design when incorporated into early design stages. Considering SA during the design stage ensures that the resulting support systems align with users expectations, as well as those of the system developers' (Endsley & Hoffman, 2002).

Mental Models, SA, and Decision Support Design. Addressing these challenges may be difficult due to the information-dense nature of many meteorological datasets; however, research into the role of mental models, sensemaking, and workload on SA and decision making help to shed light on ways to accomplish these goals (Trafton & Hoffman, 2007). Extensive work into understanding mental models has revealed that forecasters create and apply mental models when trying to comprehend the weather and project potential threats in the future (Pliske et al., 2004; Smallman & Hegarty, 2007; Trafton, 2004; Trafton et al., 2000). The need for weather forecast decision aids that support SA and decision making is well-recognized. Trafton and Hoffman (2007) call for "innovation and revolutionary redesign, especially in the creation of systems that support the forecaster in creating a graphical depiction of their own mental model."

A central question in decision support design is how to create visual displays that convey highly detailed data to users in a way that allows them to make inferences and make sense of the situation (Trafton & Hoffman, 2007). While much meteorological data is quantitative, Trafton et al. (2000) observed that forecasters primarily communicated their understanding of the data in qualitative means.

Questioning how expert forecasters comprehend and understand meteorological visualizations, Trafton et al. (2000) found that forecasters went through a process involving situational initialization, mental model building, and verification and altering of the original mental model. During initialization, forecasters primarily gathered information from discrete visualizations; the authors believed this revealed that forecasters were situating themselves with respect to the environmental context. Establishing the context did not involve a high degree of information integration and comparison; in fact, comparison characterized the mental model building phase. While building their qualitative mental models, forecasters assessed data displayed in a variety of meteorological visualizations, often comparing between datasets in a goal-oriented manner.

Comparison among information sources in order to extract information is a core activity that forecasters engage in while developing an awareness of the situation; this behavior is governed primarily by a forecaster's goals (Kirschenbaum, 2004; Trafton et al., 2000). In practice, this suggests that forecasters might benefit from decision aids that facilitate comparison and making inferences from the data (Trafton, 2004). Trafton et al. (2000) recommend that meteorological decision support systems are developed to enable comparison and integration through means such as data overlays or multi-panel displays. Additionally, Kirschenbaum (2004) found that forecasters using dual-monitor displays made more comparisons between visualizations than forecasters using a single-monitor workstation; while both groups used comparison activities to make sense of the situation, it is possible that single-monitor workstations increased the time required to understand the data, resulting in fewer total comparisons.

Minimizing Decision Bias Through Design. Facilitating data comparison is important in decision support design not only because of its frequency in the mental model building phase, but also because of its role in allowing forecasters to validate and change their mental models (Kirschenbaum, 2004). Whereas the initialization phase may create an anchor for the qualitative mental model, comparisons are then used to assess the level of fit between the anchor and additional data sources (Trafton et al., 2000). This complements the anchoring-and-adjustment model of belief updating, first proposed by Hogarth and Einhorn (1992). The anchoring-and-adjustment model posits that over time, an individual first develops a belief about a situation based on initial exposure to information. Through exposure to new data, this initial anchor may be adjusted or confirmed through a variety of processing mechanisms (Hogarth & Einhorn, 1992). Wickens et al. (2008) suggest that attentional capacity plays a large role in belief updating and situation assessment; when a user is able to devote adequate attention to data, he or she may be able to achieve a higher level of SA. In terms of meteorological decision support design, this implies that systems that direct attention to important components of the data may allow forecasters to develop high SA.

While this behavior is a critical part of the situation assessment process, heuristics such as anchoring and representativeness may bias a forecaster's judgment (Doswell, 2004). Representativeness refers to the level of similarity between two situations, whereas anchoring refers to the action of locking into a base state of knowledge (Tversky & Kahneman, 1974). In weather forecasting, representativeness bias displays itself when a forecaster misidentifies a weather event based on a perceived similarity between the event in question and a prototypical event (Doswell, 2004).

Particularly in situations with great uncertainty, anchoring bias may diminish a forecaster's ability to develop and maintain SA if an appropriate update to SA has not occurred (Nadav-Greenberg, Joslyn, & Taing, 2008).

Fortunately, certain design characteristics of visualizations may reduce the effects of these decision biases. Nadav-Greenberg et al. (2008) evaluated professional forecaster performance when using three types of uncertainty visualizations to determine wind speed forecasts: a box plot chart, a margin-of-error chart, and upper bound visualization. While a box plot chart was the most readable of the three—subjects were able to interpret the range of possible wind speeds quickly—a margin-of-error chart provided the most effective in situation assessment of uncertainty. Furthermore, Nadav-Greenberg et al. (2008) found that as forecasters became more aware of the uncertainty in the data, the lower their confidence was in their forecasts. This finding has implications for the design of decision aids that support the situation assessment cycle in uncertain situations such as weather forecasting. Interpretation of meteorological data and uncertainty information can also be conveyed by incorporating knowledge from the vast collection of color scale research (Hoffman et al., 1993). Appropriate color usage in meteorological visualizations can support decision making by drawing the user's attention to critical areas of the data and may improve task completion time (Trafton & Hoffman, 2007).

Based on the findings of Stewart, Heideman, Moninger, and Reagan-Cirincione (1992), it can be argued that weather forecasting decision support systems should adequately balance information quality with quantity. In a series of studies to assess the effects of information quality and quantity on forecast skill, forecasters produced short-term severe weather forecasts while exposed to a variety of information conditions. Stewart et al. (1992) found that as information quantity increased, disagreement between forecasters also occurred more frequently. In addition, forecast accuracy improved as the quality of the provided information improved, though accuracy did not significantly improve as quantity increased. This was especially observed in scenarios where the provided data contained great uncertainty, and in combination with the findings of Nadav-Greenberg et al. (2008) supports the notion of designing decision support that incorporates uncertainty estimates.

Automated Decision Support for Forecasting. Trafton (2004) argues that meteorological visualizations should focus on qualitative aspects of the data, opposed to only displaying quantitative information, in order to facilitate mental model building and decision making. Likewise, Hoffman and Coffey (2004) suggest that decision support design should be concerned with assisting forecasters with "generating, manipulating, and verifying a graphical 4-D representation of their mental models of atmospheric dynamics." Automation has shown promise as a means to accomplish this, and can potentially be used to facilitate SA development (Dao et al., 2009), workload reduction (Karstens et al., 2015), and comparison between data sources (Trafton et al., 2000).

In the weather forecasting domain, automation could assist forecasters in verifying and updating their SA by presenting users with pre-selected collections of visualizations (Trafton et al., 2000). Hypothetically, this may reduce workload and lead time as it would lessen the effort and time required in order to select and display visualizations. Decision support systems could be automated to display different visualizations at the most relevant points in the forecast process (Trafton et al., 2000). It is well established that forecasters consult different types of visualizations at different points along the forecast timeline (Morss & Ralph, 2007; Trafton et al., 2000), so an automated display mechanism could be an effective means of streamlining the situation assessment process.

Empirical evidence exists in support of using animation to expedite situation assessment in novice forecasters. Lowe (2004) found that among novice users of meteorological visualizations, animating changes in the geospatial data over time did not result in improved comparisons or forecasting performance. Thus, it is recommended that in order to improve novice user performance, dynamic visualizations ought to contain supplemental information to guide users to areas most relevant to the user's goals (Lowe, 2004, 2008). Although these findings were observed in novice meteorologists, such supplements could perhaps aid expert forecasters building mental models when unfamiliar situations are encountered.

In the context of river flood forecasting, Pagano et al. (2014) points out that although some scholars advocate fully automated decision support systems, keeping the operator in-the-loop with interactive automations adds value to forecasts. Recent developments in meteorological decision aiding automation have had generally positive outcomes. A particular type of automated guide, the recommender, synthesizes information from multiple sources and generates a suggestion to the user regarding locations at risk for a particular type of environmental threat. Although their purpose is to reduce forecaster workload while increasing lead time and forecast accuracy, Karstens et al. (2015) failed to find a significant difference in forecast creation time based on presence and absence of a recommender for severe hail. However, recommenders may have a greater effect on SA. In an evaluation of an automated decision aid for air traffic management, Dao et al. (2009) found that level of automation produced a significant effect on SA as measured by response time. Subjects performed a conflict resolution task with varying levels of interaction with the decision support automation. Subjects exhibited higher SA when under the interactive decision support condition than when under the fully automated condition, in which decisions had to align with the automated recommendation. This aligns with Endsley and Kiris (1995), who also found that decision makers experienced diminished SA and performance after exposure to a fully automated decision task. However, an appropriate balance between human analysis and automation use may lead to improved SA.

Summary of the Weather Forecasting Literature

Despite incorporating many technological systems for data analysis, forecasting is an inherently human activity. As a complex sociotechnical system, the weather forecasting domain involves a vast number of information transfers between human operators and non-human information sources, such as visualizations representing numerical weather prediction models and environmental observations. Recommenders and other automated decision support systems have the potential to revolutionize the situation assessment process, while understanding forecaster mental model development reveals connections between cognition and decision support technology. Furthermore, designing systems to support SA and decision making should lead to improved forecast accuracy and lead time while reducing forecaster workload. Though many advances have been made to decision support systems for weather forecasting, there remain many questions related to the relationship between human cognition, SA, and human-system integration.

In the broader picture of SA and decision making, a number of gaps exist in knowledge related to the role of cognitive processes, uncertainty, and human performance in weather forecasting emerge. Although models of sensemaking like the Data/Frame Theory in part explain aspects of uncertainty management in decision making (Klein et al., 2007), the widely-accepted models of individual SA fail to explain the role of imperfect information in situation assessment. According to the Endsley 1995 Model of SA, errors occur due to misperception, incorrect mental model choice, working memory limitations, and attentional capacity limitations (Endsley, 1995c). However, these types of errors assume that the subject has a low level of SA, and does not explore errors that occur when a subject is highly aware but immersed in an uncertain situation. This scenario may occur in weather forecasting when there is high uncertainty involved in predicting environmental conditions. Weather forecasters often consult specialized guidance products aimed at developing their awareness of uncertainty within relevant data sources, and such information can even assist with the forecast verification process, a critical component of situation assessment (Novak et al., 2008).

Doswell (2004) calls for increased collaboration between decision-making scholars and weather forecasting scholars; such interdisciplinary work is necessary to synthesize methods and theories from multiple fields with the ultimate goal of creating a forecasting system that complements human decision making processes. Theories of SA and sensemaking processes may illuminate less understood areas of cognition in weather forecasting, such as how weather forecasters use mental models to manage uncertainty at different points along the forecast timeline (Pliske et al., 2004). Furthermore, human factors research can be used to explore the relationship between complex visualization design and expert decision making under uncertainty. The weather forecasting domain offers a unique application for exploring the role of uncertainty in situation assessment, and such research has great potential to result in real and meaningful societal impacts.

Chapter 3: The Effect of Display Design on Situation Awareness in Flash Flood Detection

Submitted to the International Journal of Human-Computer Studies (Under Review)

Introduction

In the field of weather forecasting, computational modelers are under pressure to provide actionable information to forecast consumers at increasingly local levels, pushing gridded forecasting systems to hyper-resolution scales (Wood, et al., 2011; Beven, Cloke, Pappenberger, Lamb, & Hunter, 2015). Although the capability to predict weather phenomena at small scales continues to develop, operational technology often limits display capacity. Tools such as large high-resolution displays have been shown to overcome the data abstraction limits while enabling users to engage in exploratory data analysis (Lehmann, Schumann, Staadt, & Tominski, 2011); however, current operational forecasting displays are based on the multi-screen desktop setup, and meteorological visualization environments are constrained to comparatively low resolution displays.

One such set of gridded forecasting product is the Flooded Locations and Simulated Hydrographs (FLASH) project for flash flooding prediction. In July 2013, the Hydrometeorological Testbed at the Weather Prediction Center (HMT-WPC) hosted the first Flash Flooding and Intense Rainfall (FFaIR) experiment (Barthold et al., 2015). The purpose of the experiment was to evaluate the utility of a set of experimental forecast models, including the FLASH Return Period visualization, on a sample of professional forecasters and weather researchers. Over the three-week period, forecasters assessed operational and experimental computational models to create probabilistic forecasts of heavy rainfall and flash flooding events in the United States.

As part of the experiment, the researchers observed forecaster behavior when creating the forecasts and identified patterns of information processing. Through daily observations of three independently acting forecasters, the researchers observed that the design of the information display affected how well forecasters were able to interpret the data modeled in FLASH.

Of particular interest was a forecaster belief that the sampling and aggregation methods employed in the FLASH Return Period visualization led to an increased number of false alarm forecasts. The underlying grid for all FLASH visualizations covers a spatial extent of the continental United States at a horizontal resolution of 1 km. In the Return Period visualization, the experimental model calculates a measure of flash flood risk, the return period[1], for every cell within the grid; this calculation is based on a hydrologic model. However, when fully zoomed out to show the map of the entire continental United States, desktop-based display systems are not able to display each individual grid cell. This issue was overcome by developing an aggregation algorithm to sample the maximum grid cell value out of a collection of at least 112 grid cells contained within one pixel, and the map of all the maximum values displays at the national level. In practice, this means that while the true predicted return period values are displayed when a viewer zooms in to a local level, the national view displays an adjusted value of the data by displaying the maximum value. An example of this phenomenon is shown in Figure 7. At the national level, this resulted in an occlusion effect, where lower return period values were occluded by the maximum values.

[1] A return period is a measure of likelihood of some event occurring. In hydrologic terms, a return period is the average length of time for a certain threshold of flooding to be reached (Mays, 2010).

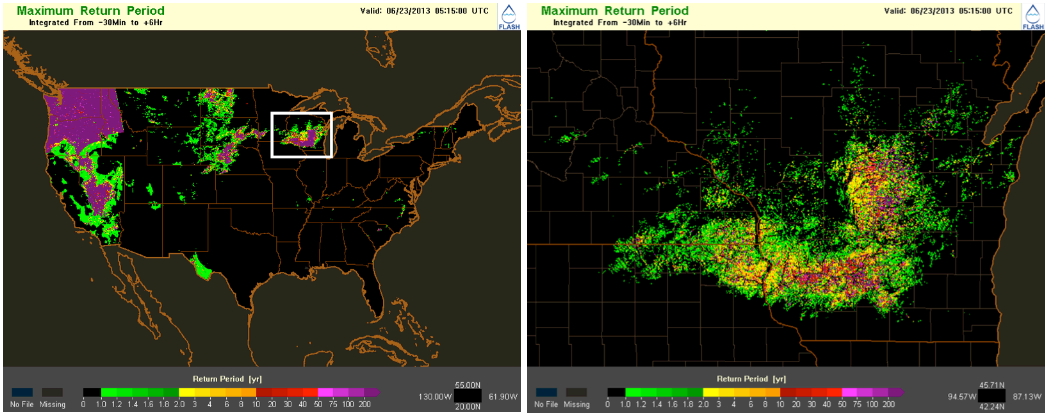

Figure 7. National view component of a stimulus sequence in the maximum sampling algorithm condition (on left) and the corresponding local view, selected within the white box (on right)

In terms of cognition and geospatial data, the field of cartographic communication has done much to inform the field of meteorological visualization and decision-making. Cartographic researchers have long studied issues related to visualizing spatial data, and have addressed such issues in terms of design and user-centered evaluation (Dobson, 1979; Hegarty, Smallman, & Stull, 2012; Montello & Freundschuh, 2005). One challenge encountered in designing geospatial visualizations is that of reducing selection occlusion (Shrestha, Zhu, & Miller, 2014). In designing visualizations that include data aggregates, Elmqvist and Fekete (2010) recommend following the principles of visual summary and awareness of fidelity. The principle of visual summary states that the visual properties of the data aggregate should be representative of the individual data point members; however, certain aggregation methods can lead to loss of fidelity and misinterpretations of the visualization. Elmqvist and Fekete (2010) point to inadequacies involved in using average-based aggregation methods, due to the loss of knowledge about the variance within the aggregate; to overcome issues related to fidelity, they recommend the use of interactive visualization overviews. The concept of overview in information visualization has been discussed extensively in the literature, with a now well-known keystone in Shneiderman's (1996) Visual Information Seeking Mantra: "overview first, zoom and filter, then details on demand." Hornbæk and Hertzum (2011) pose two alternate views on the meaning of "overview" based on a comprehensive literature review.

The core contribution from this research relates to understanding the relationship between visual data aggregation and weather forecasters' situation awareness. As defined by Endsley (1995c), situation awareness is the ability to perceive elements within a system, comprehend their significance, and project their meaning into the future in order to make a decision. In theory, strong SA should translate into the ability to make informed decisions (Adams, Tenney, & Pew, 1995). However, acquisition of SA is not discrete, but develops over time as decision-makers gain experience with and exposure to their operating environment (Endsley, 2015b). Underlying the concept of SA are a variety of personal factors and cognitive mechanisms, including, but not limited to, visual information processing, cue detection, working memory, goals, preconceptions, background training, and system design (Endsley, 1995c, 2015; Hoffman, 2015).

The Research Question

Forecaster comments from FFaIR led the researchers to hypothesize that a display algorithm that takes the average of sampled grid cells (henceforth called the average-based display) would produce different task performance than the maximum-based display. Using an empirical approach, the present study identified differences in terms of error rate and task completion time when comparing two different display algorithms on the national-scale maps. The research question addressed in the following chapter is:

RQ1: How does data aggregation in a FLASH visualization affect user performance in terms of signal detection, task completion time, and congruence in decisions for a flash flood prediction task?

Hypotheses

It was hypothesized that the aggregation method would affect false alarm rates (i.e. the event was forecast but not observed) and hit rates (i.e. the event was correctly forecast and observed). In terms of task completion time, it was hypothesized that the aggregation method would also affect the time it took participants to evaluate the displays. It was thought that due to the color scheme and the larger area of represented regions, the design of the maximum-based display would draw attention to events more rapidly than the average-based display would. Formally, the hypotheses made in the following chapter are:

H1.1: Hit Rate (HR) and False Alarm Rate (FAR)

$H_0 : HR_{avg} = HR_{max,} \qquad FAR_{avg} = FAR_{max}$

$H_1 : HR_{avg} \neq HR_{max,} \qquad FAR_{avg} \neq FAR_{max}$

H1.2: Task Completion Time (t)

$H_0: t_{avg} = t_{max} \qquad H_1: t_{avg} \neq t_{max}$

H1.3: Congruence between views (C)

$H_0: C_{avg} = C_{max} \qquad H_1: C_{avg} \neq C_{max}$

Method

Experimental Design

As a between-subjects independent variable, the display algorithm differed across two levels—participants viewed either the maximum-based display or the average-based display. Though property damage was used as a measure of severity to select the images, it was not an independent variable itself—within the study framework, participants had to detect which images represented severe events. Likewise, while participants viewed images at the two spatial scales, the local images were identical no matter which display algorithm each participant viewed at the national level. The purpose of viewing identical local images was to identify whether or not there was any bias in detection based on which level of national image a participant viewed first.

Using a Signal Detection Theory framework, error rates were calculated from the response data from the detection task (McNicol, 2005). In traditional explanations of error rate analysis in weather forecasting, signal detection metrics are based on comparisons between the predictions and the actual outcomes. For example, a hit would occur when a flash flood was forecast and then actually occurred. A false alarm refers to an event in which a flash flood was forecast but then did not occur. Translated into the present study's framework and shown in Table 1, in which all stimuli visualized

Table 1. Interpretation of error rates in the property damage detection task

		Reports of Property Damage	
		High Value	Low Value
Forecast of Property Damage	High Value	Hit	False Alarm
	Low Value	Miss	Correct Rejection

flash floods that received reports, the explanations of error rates is instead based on correct identification of property damage level for NWS-verified flash floods.

Materials and Equipment

A set of 40 image sequences was created by taking screen captures of FLASH. Each sequence consisted of one image of FLASH at a national, full-view level, and a second image of the same date and time, but zoomed in to a local level covering several counties. It is important to note that while participants in the two display groups viewed different representations of the weather event at the national scale, the local images that participants viewed were identical between groups. An example of an image sequence using both display conditions is shown in Figure 8.

The dates and times were selected based on flash flooding events that were reported between April and July 2013 in the National Climatic Data Center Storm

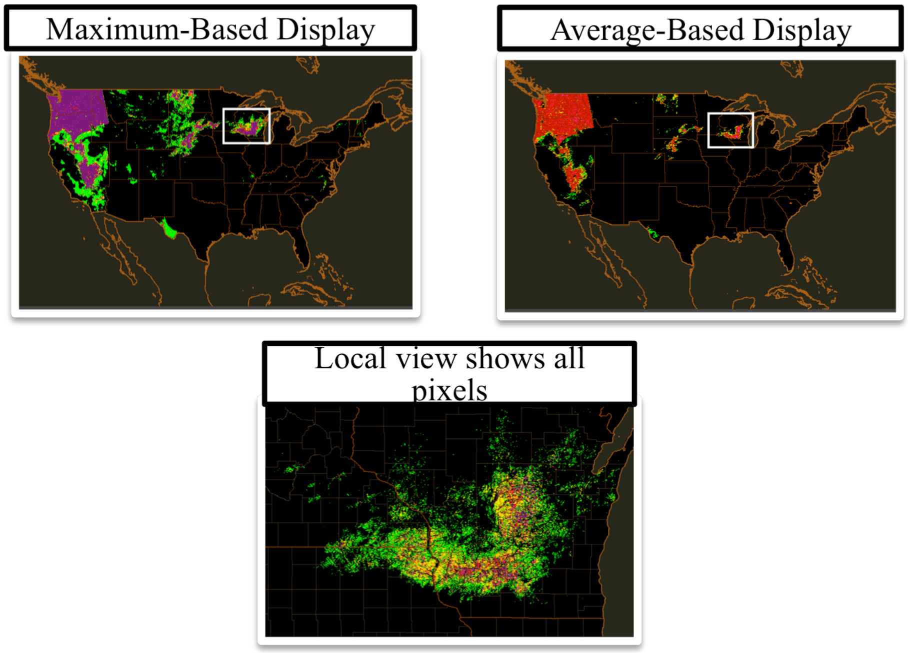

Figure 8. Examples of image sequences in both aggregation conditions, with the corresponding local-level image

Events Database (National Climatic Data Center, 2014). When selecting the events from the database, the researcher categorized events into “severe” (high value) and “not severe” (low value) flash flooding. Unlike tornado events and the Fujita scale, there is not yet a standardized scale for flash flooding severity, so the research team defined severe flash flooding to be those that caused $500,000 or more of property and crop damage ($n = 20$, $\mu = \$9.86M$; $\sigma = \$22.33M$). Events that were placed in the “not severe” category had less than $500,000 of property and crop damage ($n = 20$, $\mu = \$38.75K$; $\sigma = \$84.59K$). This distinction was explained to participants prior to beginning the study.

It is important to note that all stimuli contained models of rainfall events that were associated with NWS-verified reports of flash flooding. Although selecting stimuli based on presence and absence of flash flooding was considered, this design was determined to be too subjective. Particularly in rural, unpopulated regions, lack of an NWS-verified flash flooding report is not evidence that a flood did not occur; if it is deemed unlikely that any people or property were affected, a report is not always made to an official record. Thus, the scope of this research extends only to events connected to NWS reports.

Images were randomly presented to participants using PsychoPy, an open-source software that allows researchers to present stimuli and collect response data from participants (Peirce, 2007). Each evaluation was conducted on an Asus A53U laptop with a 15-inch screen; each image was displayed at a size of 869x680 pixels.

Participants

The sample consisted of 30 participants recruited from the student and post-doctoral population at the University of Oklahoma. Participants were required to either be pursuing a degree in meteorology or atmospheric science or to already possess one. This expectation ensured that they had adequate experience with reading weather prediction visualizations. However, participants had little experience working with the FLASH system. In terms of gender and age, the participant pool included 19 males and 11 females between the ages of 21-41 years old, with a mean age of 25.0 years and median age of 23 years. Participants were randomly assigned to one of the between-subjects display conditions (the maximum-based algorithm or the average-based algorithm).

Procedure

Initially, participants were informed about the study's purpose and tasks. After completing an informed consent form, participants received an excerpt from the FLASH training manual that explained how to read and interpret the FLASH display with pictorial examples. During the instruction stage, participants were given the opportunity to ask questions about FLASH, how to interpret the display, and what the study would involve.

Once participants stated that they felt comfortable with the FLASH interface, they answered a series of demographic questions (age, gender, and academic classification). Following this, participants viewed the image pairs presented in a randomized order. In terms of signal detection theory, the goal was to detect a high threat level (the signal) from the noise (a low threat level). In each sequence, the first image showed an event in FLASH on the national-level. Participants were asked, "Based on the information that is modeled in this image, would you expect for this event to produce flash flooding with severe levels of property damage? (>$500,000)." Participants reviewed the image, and then pressed "y" for yes or "n" for no after making their decision. The following image always represented the same weather event, but visualized at the local scale. The participants answered the same question about severity based on the new presentation. The forty image pairs were presented in a randomized order. When participants finished with the final pair, they were debriefed.

Results

Error Rates

After collecting the participants' responses, the error rates in terms of the Signal Detection Theory framework were calculated for the severity judgment associated with the average-based and maximum-based display styles and for the national and local images (McNicol, 2005). The data were compared using t-tests. A summary of the results is shown in Table 2. The results show that there is a significant difference between the display methods. The maximum display produced a higher hit rate than the average display ($p < 0.0001$), but the average display minimized false alarms ($p < 0.0001$).

A similar analysis of participant judgments was done for the local-level images. Though all subjects saw the same images in this category, responses were compared between the maximum-based and the average-based participant groups in order to ensure parity. As expected, a t-test found no significant difference between how participants in either test group when judging the local-level images. Still, as shown in

Table 3, participants did not make perfect judgments, which may in part be due to lack of participant experience with flash flood forecasting.

Bias and Sensitivity

A sensitivity index (d’) was calculated for both the average-based display and the maximum-based display. The d’ scores for the average-based display and maximum-based display were 1.00 and 0.93, respectively. This indicates that there is little difference between the discriminability of a severe flood signal between the two display types. In addition, a significant difference was found in the biases associated with the two display algorithms ($p < 0.001$): for the maximum-based display algorithm, a liberal bias of -0.74 was found, and a conservative bias of 0.24 was found for the average-based display algorithm. This can be interpreted to mean that participants in the maximum-based display condition were more likely to say that any stimulus contained a significant flood, while the participants in the average-based display condition were more likely to say the opposite.

Table 2. Comparison of average-based and maximum-based display types in terms error rates

	Hit Rate	**False Alarm Rate**	**Sensitivity *(d')***	**Bias**
Average Algorithm	0.57	0.25	0.93	0.24
Maximum Algorithm	0.85	0.50	1.00	-0.74
p-value	<0.0001	<0.0001		

Table 3. Error rates for viewing the local-level events (n = 30)

	Hit Rate	False Alarm Rate
Local	0.50	0.24

Task Completion Time

Task time was recorded from the time the national display was shown to the participant until they made a severity judgment on the national image. A summary of results is shown in Table 4. Though the original hypothesis was that aggregation method would affect response times, no significant differences were found between the two. A further analysis of task completion time for the time taken during hits, misses, false alarms, and correct rejections also failed to find any statistically significant differences (hit: p = 0.58; miss: p = 0.81; correct rejection: p = 0.12; false alarm: p = 0.57).

Effect of Display Design on Congruent Decisions

Congruent decisions, or those in which the response for the national image was identical to the response for the corresponding local image, were measured between display conditions. Congruent decisions were deemed either *congruent-correct* (a "yes/yes" response to an image sequence that represented a significant flood or a "no/no" response to an sequence that represented an insignificant flood), *congruent-incorrect* (a "no/no" response to a sequence that represented a significant flood or a

Table 4. Average time (in seconds) taken to produce a hit, miss, correct rejection, or false alarm, analyzed with a t-test.

	Hit	Miss	Correct Rejection	False Alarm
Average Algorithm	4.02	4.49	4.34	4.12
Maximum Algorithm	4.63	5.23	5.81	5.45
p-value	0.58	0.81	0.12	0.57

"yes/yes" response to an image sequence that represented an insignificant flood), or *incongruent* (a "yes/no" or "no/yes" response, which by definition was always partially correct). Counts of congruent and incongruent decisions by display condition are shown in Tables 5 and 6.

A Chi-squared test of decision counts against display condition revealed a significant difference between the maximum sampling display algorithm and the average sampling display algorithm for judgment congruence; however, these differences were observed when assessing judgment congruence in relation to threat level. When judging images representing low property damage event, participants in

Table 5. Counts of congruent and incongruent decisions by display condition for high-level threats

High Property Damage (Threat Level)				
	Hit/Hit	Miss/Miss	Hit/Miss or Miss/Hit	Row Totals
Max. Display	148 (52.0%)	40 (14.0%)	97 (34.0%)	285 (100.0%)
Avg. Display	99 (33.0%)	84 (28.0%)	117 (39.0%)	300 (100.0%)
p-value		< 0.0001		

Table 6. Counts of congruent and incongruent decisions by display condition for low-level threats

Low Property Damage (Threat Level)				
	Hit/Hit	Miss/Miss	Hit/Miss or Miss/Hit	Row Totals
Max. Display	139 (46.3%)	66 (22.0%)	95 (31.7%)	300 (100.0%)
Avg. Display	204 (68.0%)	48 (16.0%)	48 (16.0%)	300 (100.0%)
p-value		< 0.0001		

the average display condition produced more congruent judgments (hits on both images within a given image sequence) than participants in the maximum display condition, χ^2 (2, N = 600) = 31.16, p < 0.0001. Conversely, when judging an image representing a high property damage event, participants in the maximum-based display condition produced more congruent hits and fewer congruent misses than those in the average-based display condition, χ^2 (2, N = 585) = 36.15, p < 0.0001.

A closer look at the data indicates that the threat level factor may have also affected congruent choices. As shown in Table 7, regardless of the display condition, participants made significantly more correct congruent hit judgments when viewing low property damage events and were more likely to congruently miss an event that was a high threat level, χ^2 (2, N = 1185) = 29.98, p < 0.0001.

The Relationship Between Event Size and Response

As the findings from the aforementioned analyses suggest that display condition and threat level did impact decision accuracy, a random-intercept logistic regression was selected to estimate the likelihood of producing a correct response given certain conditions. Due to the binary nature of the responses (0 = incorrect, 1 = correct), a logistic regression was chosen as the appropriate method to determine the relationship

Table 7. Counts of congruent choices by threat level, independent of display type

	Hit/Hit	Miss/Miss	Hit/Miss or Miss/Hit	Row Totals
High Threat Level	247 (42.2%)	214 (36.6%)	124 (21.2%)	585 (100.0%)
Low Threat Level	343 (57.2%)	143 (23.8%)	114 (19.0%)	600 (100.0%)
p-value		< 0.0001		

between event size and response. The random-intercept logistic regression accounts for interdependencies among repeated observations within subjects and adds a subject-specific random intercept to the regression equation (Rabe-Hesketh & Skrondal, 2008). Using this type of regression, we were able to draw conclusions about odds ratios adjusted for individual differences; these are sometimes called subject-specific probabilities. The resulting logistic regression equation is shown in Equation 1.

$$\begin{aligned} logit\big(p(x)\big) &= \log\left(\frac{p(x)}{1-p(x)}\right) \\ &= 1.87 - 4.51x_{size} - 1.93x_{threat} - 1.23x_{display} \\ &+ 5.33x_{size}x_{threat} + 2.86x_{size}x_{display} \\ &+ 2.48x_{threat}x_{display} - 3.15x_{size}x_{threat}x_{display} \end{aligned} \qquad (1)$$

Initially, the model included two main effects, *display type* (d) and *threat level* (t). However, prior work has suggested that task-irrelevant features on geospatial displays may negatively impact task performance (Hegarty et al., 2012). Although participants were instructed to judge only the area within the white selection box on each stimulus, many of the stimuli contained visually distracting imagery of flood predictions outside the box. Thus, after the data collection phase, each stimulus received a code to designate the amount of visual distraction as determined by geographic scale of the area mapped with return period values; the new explanatory variable, *size* (s), was created with two levels: *small* and *large*. An example of a small-scale stimulus and a large-scale stimulus are shown in Figures 9 and 10, respectively.

Interestingly, the random-intercept logistic regression produced an estimated variance between subjects of zero ($\psi \approx 0$), which led the random-intercept model's explanatory variable coefficients to converge with those of the ordinary logistic regression model.

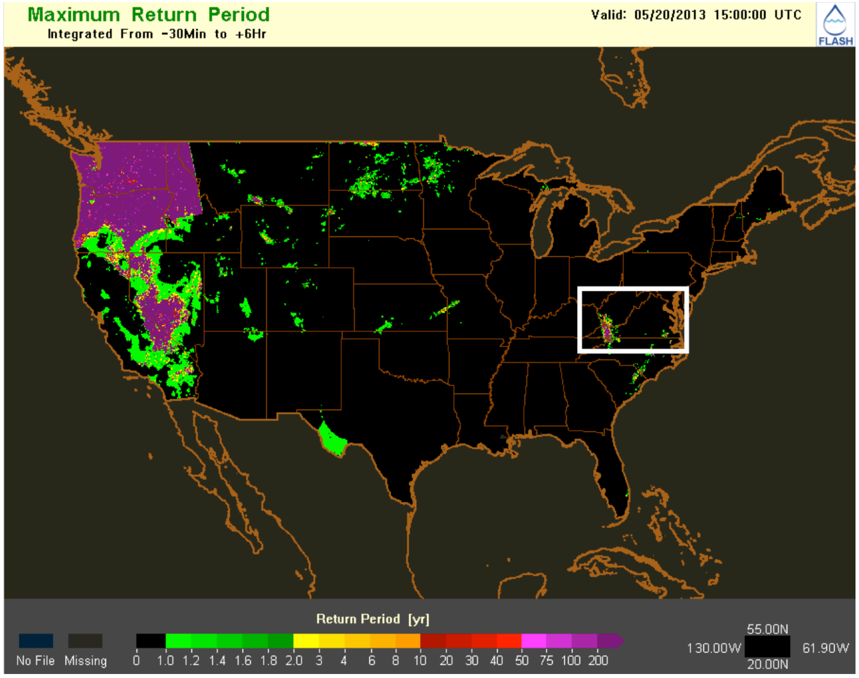

Figure 9. National-scale stimulus where size = small, threat level = insignificant (< $500,000)

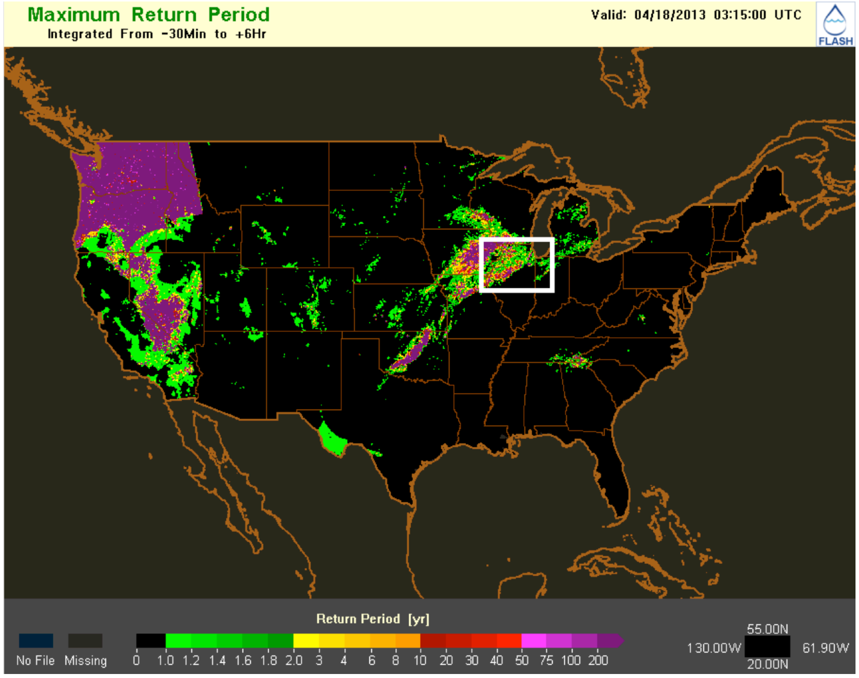

Figure 10. National-scale stimulus where size = large, threat level = significant (> $500,000)

A comparison of the ordinary logistic regression and random-intercept logistic regression is presented in Appendix A.

In logistic regression, the odds ratio for each term reflects the ratio of the odds between giving a correct response at the x=1 level and the x=0 level. For example, the odds ratio for the size variable is 0.011 (with a 95% confidence interval of 0.003 – 0.032); this is interpreted to mean that the odds of a participant giving a correct response to a stimulus the contained a large-scale event were 0.0110 times the odds of giving a correct response when viewing a stimulus that contained a small-scale event. In reverse, the odds of participant being correct when viewing a small-scale event were approximately 90.909 times the odds of correctly judging a large-scale image. Likewise, the odds of a participant correctly judging a stimulus image that contained a significant threat of property damage were 0.145 times the odds of correctly judging an image with insignificant levels of property damage (with a 95% confidence interval of 0.089 – 0.230). Finally, the odds of a participant correctly judging a stimulus when visualized with the maximum-based display algorithm were 0.293 times the odds of correctly judging a stimulus displayed with the average-based algorithm. All two-way and the three-way interaction between explanatory variables were significant, and the associated odds ratios are shown in Appendix A.

The Likelihood Ratio Test was used to evaluate model fit between the full and the intercept-only model ($H_0\colon \beta_{0,full\,model} = \beta_{0,reduced\,model}$), which produced a significant p-value (p < 0.0001). Thus, the full model was selected as the model with better fit to the data. Further analyses which compared the full model to single-term models had statistically significant p-values from the Likelihood Ratio Tests, so the single term models were deemed to be too simplistic to adequately characterize the relationship.

The Hosmer-Lemeshow test was used to assess the model's goodness of fit. The Hosmer-Lemeshow test is a significance test, similar to the Chi-Square test, and compares the variance of the specified model to the variance of the null model. Some limitations to consider involve sample size: the larger the sample size, the poorer the test's performance. Another limitation is common to other significance-based goodness of fit tests: it only indicates whether or not the model fits better than the null model, and includes no indicator of how well it fits the data. In the present study's case, the Hosmer-Lemeshow test's p-value was approximately 1, which fails to reject the null hypothesis. This indicates that full model and the null model have equal variances, and that the full model adequately fit the dataset. However, we must consult alternative measures to determine the level of fit, as the test does not reveal the degree of the model's fit.

For a logistic regression, two variations of Pseudo R^2 terms are used to assess a model's goodness of fit. The first category of Pseudo R^2 assesses model fit in terms of a specified model's improvement over a null model; this is similar to the approach taken by the Hosmer-Lemeshow test. The Cox and Snell Pseudo R^2 as well as Nagelkerke's Pseudo R^2 take this approach, using null and specified model log likelihood values to calculate Pseudo R^2. Nagelkerke's Pseudo R^2 tends to be more intuitive to interpret than Cox and Snell's value; the Nagelkerke Pseudo R^2 value is essentially a standardized version of Cox and Snell's value and is measured between 0 and 1 (unlike

Cox and Snell). In terms of the full model in this test, Nagelkerke's Pseudo R^2 value was 0.275.

The second type of Pseudo R^2 estimate interprets R^2 as a metric to explain variance within a model using residuals to measure variability. Examples of Pseudo R^2s that take this approach include McFadden's Pseudo R^2, Effron's Pseudo R^2, and McKelvey and Zavoina's Pseudo R^2. In terms of McFadden's Pseudo R^2, a value between 0.20-0.40 is considered to fit the model exceptionally well (Louviere, Hensher, & Swait, 2000).

In the context of the present tests, McFadden's Pseudo R^2 was 0.173. While the Pseudo R^2 values indicate that the specified model is only a moderate improvement over the null model, the Hosmer-Lemeshow test affirms that the model fits the data set. However, while the sample is not extraordinarily large (n=1185), the Hosmer-Lemeshow test is sensitive to large sample sizes, so it is possible that the sample size has affected the outcome of the test.

Discussion

The results show that choice of display method did influence probability of detection and false alarm rate. Furthermore, the hypothesis that the aggregation method would affect the hit rates and false alarm rates was supported. However, no difference in task time was found between display methods. The logistic regression analysis revealed that while the display condition did affect the likelihood of a correct response, the predictors of threat level and visual size, as well as all interactions, were also significant explanatory variables for judgments of the national-level stimuli. When evaluating the likelihood of producing a correct response on the local-level stimuli, the logistic regression analysis also showed that correctness on a national-level stimulus was a significant predictor on producing a correct response for the corresponding local-level stimulus. Likewise, this relationship was explored in the congruent-choice analysis, which indicated that threat level and display condition were likely to affect congruence in decision-making between the national- and local-level stimuli.

Data Aggregation in Visual Decision Aids

The analysis of congruent-correct and incongruent decisions for threat level and display condition showed that the average-based sampling display algorithm was a poor aggregation technique: when visualizing a significant threat, the average-based display led to a divergence in judgments between the national- and local-scale stimuli. The average-based display was only connected to a significant increase in congruent decisions when the stimulus contained an insignificant threat. This may be due in part to latent participant factors such as poor understanding of the return period metric. It is also possible that when a threat is minimal, the variability among data points is smaller than the variability among grid cell values for a significant threat; thus the value produced by the average-based algorithm to represent the data aggregate at the national level more closely represented the individual members within the collection. While it is debatable whether or not a correct congruent response is more desirable than an incorrect congruent response, the results show that the display condition did affect fidelity.

When speaking of visualization that incorporates data aggregation, Elmqvist and Fekete (2010) recommend that designers keep the principles of visual summary and fidelity in mind; visualizations of aggregated data ought to represent the underlying individual data points accurately and consistently. The present study's findings suggest that the average sampling display algorithm led to participants making significantly more congruent decisions than the maximum sampling display algorithm. This would indicate that the average sampling display algorithm provided a stronger representation between the aggregated, national view and the individual data points visualized in the local view. However, several visualization studies have discussed the poor ability of an average-based aggregation method to satisfy the fidelity principle. Elmqvist and Fekete (2010) point to a caution given by Andrienko and Andrienko (2006); they warn against using average-based aggregation methods due to the nature of averages flattening out variation. In the words of Andrienko and Andrienko (2006): "the mean weight of a fruit in a basket filled with apricots and one watermelon is also not a very useful aggregate characteristic."

A visual inspection of the stimuli shed light upon the connection between design and individual task performance. Unsurprisingly, an examination of the maps that were most often identified correctly and incorrectly showed that participants tended to correctly identify maps that represented the extremes of the stimuli (either huge swaths of floods or none at all) but had more difficulty when the maps were somewhere in the middle. Misidentification of stimuli was observed in both display conditions when the stimuli sets had striking differences in visual representation between the national and local levels. For example, one stimulus contained an event that looked like a very small storm when visualized with the national-level average-based algorithm, but actually had a very severe gradient after zooming closer—an indicator of flash flooding that participants were trained to seek. Participants often judged the national image to be insignificant, but changed their minds after viewing the local level.

A closer evaluation of congruent decisions by display type and threat level supports recommendations by Andrienko and Andrienko (2006) and Elmqvist and Fekete (2010). From a design perspective, decisions that are congruent between levels of geographic scale (national versus local) are highly desirable. The researcher hypothesizes that congruence indicates an adequate level of fidelity in the aggregated data visualization. While it is debatable whether or not a correct congruent response is more desirable than an incorrect congruent response, the results show that the display condition did affect fidelity. The analysis of congruent-correct and incongruent decisions for threat level and display condition showed that the average sampling display algorithm was indeed not the ideal aggregation technique: when visualizing a significant threat, the average-based display led to a divergence in judgments between the national- and local-scale stimuli. The average-based display was only connected to a significant increase in congruent decisions when the stimulus contained an insignificant threat. This may be due in part to latent participant factors such as poor understanding of the return period metric. It is also possible that when a threat is minimal, the variability in the individual data points is smaller than the variability amongst grid cell values for a significant threat, and thus the value produced by the average-based algorithm to represent the data aggregate at the national level more closely represents the individual members within the collection.

Signal Detection and Weather Forecast Decision Making

In the experimental task, participants viewed a series of stimuli and were asked to determine whether or not each displayed a significant flood threat. Although framed with Signal Detection Theory, the task involved a combination of detection and identification activities. These are fast processes, as evidenced by the task time results, and they are governed by cognitive structures such as long term memory, working memory, schema, mental models, attention, feature identification, and monitoring, among others (Adams et al., 1995; Endsley, 1995c, 2015; Hoffman, 2015; Wickens, 2015). Detection, a function of factors including but not limited to top-down processes, expectations, and background knowledge, can be mapped to Level 1 of Endsley's 1995 Model of SA, perception. Identification involves taking a detected item and evaluating its fit into a categorical grouping, and it is also affected by experience and top-down processes (Endsley, 1995c; Wickens & Carswell, 1997). Identification can be mapped to Level 2 of Endsley's 1995 Model of SA, comprehension. The third level of Endsley's 1995 Model of SA, projection, was determined to be outside the scope of the present study's goals; however, future work could extend the present study's method from a detection and identification task to a projection task in which participants would have to choose whether or not a flash flood warning would be appropriate.

Detection and Comprehension of Flash Flood Threats. The study's primary aim was to compare two data aggregation-based display methods and evaluate their effects on novice participants' forecasting speed and accuracy for a flash flood prediction task; a broader goal was to relate FLASH display design to forecasters' SA, at least in terms of detection and comprehension of signals in the visualization. In a weather forecasting-specific model reminiscent of Endsley's 1995 Model of SA, Bowden, Heinselman, Kingfield, and Thomas (2015) proposed a model of the compound warning decision process. The three-stage cyclical process involves an initial detection phase, followed by an identification phase, and completed by a reidentification phase. In a study of the effects of variable update frequency from phased-array radar data, forecasters' probability of detection (measured in terms of hits, false alarms, and misses during a simulated forecasting task) for severe hail and wind threats differed between the detection and identification phases of the compound warning decision process. Additionally, case studies of professional forecasters have also shown that the warning decision process is affected by forecaster experience and task-relevant knowledge, risk tolerance, perceptions and beliefs about environmental states, confidence, software issues, and spatial ability (Heinselman, LaDue, & Lazrus, 2012; Smallman & Hegarty, 2007). In combination with the error rate results of the present study, this indicates that in addition to display design and information bandwidth, SA in weather forecasting is governed by a variety of cognitive, individual, and technical factors.

The detection and identification tasks in the present study can be categorized within the family of cognitive integration processes. Studies of graph comprehension distinguish specific information extraction, or processes in which a user has a goal to search and find some specific attribute in a visualization, from information integration, processes in which a user may combine multiple attributes from a visualization in order to comprehend broader meanings such as trends in the data (Ratwani, Trafton, & Boehm-Davis, 2008). Due to the map-based format of many decision aids used in weather forecasting, information integration is a fundamental activity for a forecaster to be able to develop SA. Using the FLASH return period maps as an example, examining the return period value assigned to a single grid cell provides much less meaning than evaluating the overall trends and gradients over broader geographic scales. Some models of information integration for graph comprehension are represented as an iterative process consisting of pattern recognition and interpretation, in which features are detected and, ideally, understood; as graph complexity increases, more iterations of the integration process are required (Ratwani et al., 2008). Out of a vast collection of design recommendations for visual displays, several guidelines and challenges exist that may improve weather forecasting displays including improving visual discriminability (Dobson, 1979; Wickens & Carswell, 1997), highlighting meaningful information clusters to facilitate integration (Ratwani et al., 2008), and structuring the information landscape in a way that assists the user to achieve their goals in a hierarchical needs-based order (Hoffman & Woods, 2005; Trafton & Hoffman, 2007). However, while a user-centered display design can often help to overcome performance issues, performance may still suffer when users of complex displays lack appropriate background knowledge and skill (Hegarty, 2011; Shah & Freedman, 2011). While participants in the present study were not expert forecasters, their background in atmospheric and environmental science positioned them as novice system users. It is likely that a future study extending this work to expert forecasters would reveal different patterns of threat detection.

Spatial Visualizations and Response Time. In practice, the FLASH tools update dynamically, but in the experimental context, participants viewed static representations of the return period model at a single timestamp. Task times were lower than anticipated, however, this may be explained by the static nature of the stimuli in addition to experience level of the participants. Studies of cartographic interpretation have indicated that a user's background experience can affect response time during map comprehension tasks (Ooms, De Maeyer, & Fack, 2013; Ooms, De Maeyer, Fack, Van Assche, & Witlox, 2012). In an eye-tracking analysis of map-based visual search tasks, novice map users spent more time searching for specific features than expert users (Ooms et al., 2012). When map complexity was increased to include color coding and topographical detail, color codes for certain geographic features tended to attract attention away from more relevant map elements (Ooms et al., 2013). While weather forecasting typically requires more integrative processing than specific information extractions, it is possible that these findings could be extended to the present work. In the present study, participants were novices in terms of exposure to the FLASH visualization, but the requirement to be a current student or graduate of a meteorology program ensured that each participant had at least one year of exposure to weather and environmental concepts.

"Crying Wolf" in Weather Forecasting. Although the FLASH tools are intended for use by a population of professional forecasters and not members of the general public, display methods that influence false alarm rate may lead to unnecessary warnings and the "cry-wolf" effect. In the weather domain, the cry-wolf effect refers to the phenomenon wherein consumers of a weather warning fail to respond adequately after a series of false alarms, decreasing their likelihood of responding appropriately to a future true threat (LeClerc & Joslyn, 2015). In response to the concern that certain display algorithms may increase the false alarm rate, one must remember that selecting an appropriate response criterion is a function of signal probability and the costs of correct and incorrect responses (Wickens & Carswell, 1997). Thus, it is important to consider the cost/benefit relationship associated with response accuracy in weather forecasting. In weather forecasting, response criterions for warning on a severe threat are not only shaped by individual information processing of uncertain information, but also by governmental policy. As discussed by Doswell (2004), from a policy perspective, false alarms are often preferred over misses, which are traditionally held in unfavorable regard. Whereas false alarms incur costs from allocating emergency response resources and may also add to a cry-wolf effect in the long run, total failure to predict a true severe weather threat can lead to significant damage and even human fatalities when protective actions are not taken. Though the present study focused on a dichotomous choice (significant versus insignificant flooding as reflected through property damage), an extension of the work could include probabilistic forecasting. If a shift in response criterion is not a viable policy option, empirical evidence is available that suggests probabilistic risk estimates attached to severe weather warnings may reduce the cry-wolf effect (LeClerc & Joslyn, 2015).

Limitations

Several factors limit the impact of this study's results. As evidenced by the possible design biases, participant judgments may have been mislead by map appearances. Like many weather forecasting decision aids, FLASH is a simulation model and not a mapping of verified observations. However, in each experimental stimulus, we showed participants FLASH maps where flooding was confirmed after the fact and then asked participants to judge whether or not they would have expected high-impact flash flooding. Therefore, participant judgments can only be accurate as the modeling. While we attempted to filter out FLASH models of flooding events that did not appear to be accurate representations, as with any simulation model, a degree of error between the model and reality is to be expected.

In addition to modeling errors, some participants had difficulty gauging flash flooding severity. Participants were instructed to produce a yes or no judgment on whether or not they believed each of the displayed models could contain a severe flash flood. The definition of severe flash flooding as corresponding to greater than $500,000 worth of property damage was chosen arbitrarily in lieu of any other metric. A limitation of using property damage as a measure of severity was that it was difficult to estimate without a general knowledge about geographic features; for example, when unfamiliar with a certain region of the United States, participants occasionally asked whether or not there were any sizeable cities located nearby. Although we encouraged participants to use any background knowledge they might have had of weather forecasting, we also pushed them to make a decision ultimately based on how the FLASH stimuli appeared. In this regard, we tested the capability of the visualization to convey threat information and essentially tried to minimize the need for extensive meteorological or geographic knowledge. It is still possible that considering property damage level increased mental workload in some participants, but it is not apparent from the time-based results; however, while response time does sometimes reveal issues with mental workload, this is not necessarily always the case. It is also possible that participants evaluated the stimuli based on a surrogate criterion as opposed to property damage.

Finally, in the experimental design, a sample size issue may limit the generalizability of the logistic regression. In the original design, the variable of geographic scale was not included, but was assigned after the error rate analysis. Thus, sample sizes were uneven between the levels of the geographic scale.

Summary

The results of this study show that there is a significant difference between display styles in terms of error rates, but not in terms of task completion time. Though the original hypothesis was that the average display would cause participants to review the image for a longer period of time, this in fact was not observed. When examining the images that participants commonly had trouble judging correctly, common causes of confusion occurred for events that had particularly different visual representations between the national and local level. For example, one event looked like a very small storm when visualized with the national-level average-based algorithm, but appeared to be very severe when zoomed closer. Participants often judged the national image to be insignificant, but changed their minds after viewing the local level.

Design recommendations based on these results for future weather information displays must rely on the risk management values of the system designers. While the maximum display style maximized hits, it also produced many more false alarms than the average display. In weather forecasting, excess numbers of false alarms can consume valuable time that forecasters could be using to analyze true threats. However, while the average display style produced fewer false alarms, participants were much more likely to miss an event; this could also result in critical consequences.

In the case of a flash flooding prediction system such as FLASH, the recommendation from these results would be to use the maximum display algorithm. Flash flooding is by nature a rapidly occurring event that can have life-threatening consequences if not predicted with enough lead time. For such a system, having a design that promotes more hits, even at the expense of producing false alarms, would ensure that forecasters' attentions would be drawn to severe events in a timely manner.

A future study based on Naturalistic Decision Making framework (Klein, 2008) would provide knowledge on how display design using data aggregation affects the acquisition of situation awareness in real-time. Whereas the present study focused on perception and comprehension, a real-time evaluation of the display methods could help to identify connections between display design and a forecaster's ability to develop SA in a dynamic manner. Additionally, future work could address limitations of the present study. While participants all had some background in meteorology and forecasting, few had specifically studied flash flood forecasting. A similar study to the present work, but run with a sample of professional flood forecasters may supplement the present study by identifying the effects of expertise on signal detection.

Chapter 4: A Mixed Methods Approach To Understanding Situation Awareness and Uncertainty in Weather Forecasting

Introduction

In order to predict environmental threats, forecasters synthesize massive amounts of data to evaluate trends over space and time (Daipha, 2015). During this process, challenges arise from uncertainties in meteorological states (initial conditions), imperfect computational models, and individual factors (Doswell, 2004). Lipshitz and Strauss (1997) conceptualize uncertainty in decision making as "a sense of doubt that blocks or delays action." They also distinguish uncertain issues ("outcomes, situations, and alternatives") from sources of uncertainty ("incomplete information, inadequate understanding, and undifferentiated alternatives"). These classifications describe weather forecasting issues, where forecasters can misinterpret or even fail to recognize uncertainty in forecasting contexts. These assessment errors may lead to improper or inadequate use of information sources, which in turn could impact the accuracy and timeliness of a weather forecast. Such forecasting challenges may even have a negative effect on the quality of communications to forecast end users (Doswell, 2004; Novak, Bright, & Brennan, 2008).

Through surveys, National Weather Service (NWS) forecasters have indicated that receiving guidance about levels and sources of uncertainty not only adds value to forecast communications, but may also complement the situation assessment process (Novak et al., 2008). Indeed, some studies have shown that the presentation style of uncertainty information can affect a user's comprehension and performance level in a weather prediction task (Nadav-Greenberg, Joslyn, & Taing, 2008). At the decision choice stage, failure to comprehend situation-based uncertainties can lead to negative outcomes, such as forecaster hedging, defined by Murphy (1978) as "the difference between a forecaster's judgment and his forecast." It is possible that forecast guidance promoting decision making under uncertainty might reduce risk while increasing forecast accuracy or timeliness.

The current work explored the situation assessment process under uncertainty in weather forecasting. Prior explanations of situation assessment under uncertainty provide limited evidence to support application to complex decision making tasks such as weather forecasting. However, uncertainty management and SA have also been considered a form of macrocognition, a science that considers certain cognitive processes from a naturalistic decision making perspective (Klein et al., 2003; Trafton & Hoffman, 2007). Sensemaking, situation awareness, uncertainty management, problem detection, and naturalistic decision making have all been classified as forms of macrocognition (Klein et al., 2003). These approaches have led to new theories of uncertainty management, but NDM proponents suggest that empirical studies can further advance decision making research (Lipshitz, Klein, Orasanu, & Salas, 2001). In line with these macrocognitive approaches, the present study described situation assessment under uncertainty in terms of management techniques and information seeking behavior during forecasting activities.

The following chapter discusses the two-stage, mixed methods analysis of forecaster decision making and uncertainty management practices. Following a presentation of background studies, an explanation is given of the context in which the two studies occurred. The first study presents findings from a quantitative analysis of information seeking behavior during real-time forecasting exercises. The second section discusses findings from focus groups in which forecasters discussed individual strategies for managing uncertainty. Additionally, focus groups produced findings with regard to the utility of a set of proposed attributes for communicating forecast uncertainty. The chapter concludes with a general discussion of practical and theoretical implications as well as with recommendations for future research.

Uncertainty and SA in Weather Forecasting

Within the literature, several explanations of decision making under uncertainty can be found. Endsley's 1995 Model of SA included uncertainty management only in the contexts of mental model building and individual confidence (Endsley, 1995c). Endsley's (1995c) model frames situation awareness (SA) as a function of cognitive processes that include attention, perception, working memory, long term memory, automaticity, and goals. In relation to long term memory, Endsley (1995c) argues that uncertainty plays an important role in situation assessment and decision making. At the individual level, Endsley (1995c) states that uncertainty may be a source of stress, which can produce a negative effect on SA. However, these arguments were made in relation to an individual's confidence level. In weather forecasting, while confidence is a large part of the decision process, it is not the only component.

Endsley and Jones (2001) found that mental models allow decision makers to synthesize and make use of data sources. Despite contrary claims presented within the literature (Endsley, 2015b), we concur with Klein's (2015) assessment that the Data/Frame theory of sensemaking complements accounts of SA. Whereas Endsley's 1995 Model of SA incorporated situational uncertainty to a small degree, macrocognitive sensemaking studies have illuminated specific components of coping with uncertainty (Klein, Moon, & Hoffman, 2006). Endsley (2015a) identified the need for a better understanding of uncertainty management and SA, suggesting that macrocognitive approaches may improve knowledge about the relationship between SA and uncertainty in complex decision scenarios.

Due to the dynamic, uncertain conditions found in many complex sociotechnical systems, Minotra and Burns (2015) have called for research into uncertainty management tactics in such decision making settings. Naturalistic decision making studies may shed some light upon specific coping tactics; Lipshitz and Strauss (1997) identified heuristics used by decision makers in a military defense decision making study. Previous accounts of uncertainty management supported a set of techniques referred to as the R.Q.P. heuristic: in this position, decision makers cope with uncertainty through reduction, quantification of remaining uncertainty, and making decisions based upon the remainder. Lipshitz and Strauss (1997) found that decision makers concurrently engaged in situation assessment and evaluation of alternatives, adapt strategies dynamically to suit the situation. Further study challenged the generalizability of the R.Q.P. heuristic to multiple domains, which led Lipshitz and Strauss (1997) to propose a tactical framework for uncertainty management based on behavioral research; they identify reduction, acknowledgement, and suppression as primary uncertainty management methods in decision making. From their findings, Lipshitz and Strauss (1997) illustrated the R.A.W.F.S. heuristic, in which decision makers manage uncertainty through reduction techniques, assumption-based reasoning, weighing pros and cons of alternate choices, forestalling the decision, and suppressing uncertainty. These methods may also apply to decisions made in non-military contexts, and the authors recommended additional research into how tactics change across domains (Lipshitz & Strauss, 1997).

Several scholars have examined uncertainty management and decision making in weather forecasting (Daipha, 2010, 2015; Doswell, 2004; Novak et al., 2008; Stewart, Heideman, Moninger, & Reagan-Cirincione, 1992). Bowden, Heinselman, Kingfield, and Thomas (2015) framed the portion of forecasting that precedes a warning as the *compound warning decision process*, which consists of threat detection, threat identification, and reidentification as the situation dynamically changes. In practice, forecasting is a goal-directed process (Trafton et al., 2000), and comparison between data sources plays a large role in situation assessment (Kirschenbaum, 2004). Doswell (2004) suggested that forecast decisions are based on logical analysis and more flexible intuitive processes; in this account, these two processes are used to perceive and comprehend uncertainty. Likewise, in a comprehensive analysis of five years of observations in a National Weather Service Weather Forecasting Office, Daipha (2015) framed this sensemaking process with a "collage" metaphor, referring to the process of integrating numerous information sources and extracting a greater meaning. In terms of weather forecasting, Daipha (2010, 2015) posited that a collage represented "a process of assembling, appropriating, superimposing, juxtaposing, and blurring disparate pieces of information." Daipha's (2015) account aligns with Doswell's (2004) discussion of logical versus intuitive forecast decision making. It is clear that coping with uncertainty during weather forecasting involves a number of cognitive processes, and it also skirts the line between art and science.

In reference to forecasters' interpretations of weather visualizations, Trafton and Hoffman (2007) suggest that weather forecasters maintain an action cue, employ recognition-primed decision making, and engage in iterative mental model building. In their macrocognitive model of forecast decision making, SA complements mental model building, and both are products of the sensemaking cycle. Nevertheless, an understanding of how uncertainty propagates over time through the weather decision making system is still lacking. Forecasters not only incorporate uncertainty information into their mental models when assessing an evolving weather situation, but they also use their mental models to convey risk to consumers such as emergency management (Morss, Demuth, Bostrom, Lazo, & Lazrus, 2015). Decision support systems may assist weather forecasters with comprehending and using uncertainty information effectively, as well as in terms of communicating environmental threats effectively.

In order to convey uncertainty in the data, decision support systems must not only leverage visualization and interface design, but system designers must also ensure that such tools are capable of presenting the correct information to users at the moments it is needed. Trafton and Hoffman (2007) echo this sentiment and suggest incorporating automation to display visualizations at relevant moments along the forecasting timeline. One component of developing user-centered decision support systems involves identification of situation awareness requirements. SA requirements refer to information attributes and sources that are necessary for a user to accomplish his or her goals; identifying SA requirements allows system designers to satisfy user needs in order to facilitate perception, comprehension, projection, and finally decision selection; (Endsley, 1994). SA requirements are dynamic and can be ascertained through a variety of methods, including Cognitive Task Analysis (Endsley, 2001), goal-directed task analysis (Endsley, 1994; Endsley & Hoffman, 2002), Cognitive Decision Method (Hoffman, Crandall, & Shadbolt, 1998; Klein, Calderwood, & MacGregor, 1989), and Cognitive Work Analysis (McIlroy & Stanton, 2011).

SA requirements analysis and SA measurement can provide actionable feedback in the initial phases of system design (Endsley, 1995a; Endsley & Hoffman, 2002). Such methods may have utility for the design of weather forecasting decision support systems. The situation awareness-oriented design (SAOD) process incorporates findings from SA requirements analysis with interface and system design guidelines, followed by an evaluation/redesign cycle (Endsley & Hoffman, 2002). A deeper understanding of SA requirements for flash flood forecasting, as well as the cognitive processes involved with situation assessment under uncertainty, will add value to decision support systems and guidance products used during the forecast decision process.

Jones, Quoetone, Ferree, Magsig, and Bunting (2003) investigated mental simulation and pattern matching during flash flood forecast decision making. Their analysis revealed that forecasters used different information sources to guide their threat level assessments at different points in time; this difference was attributed to participant experience level, regional knowledge, and mental model and schema availability (Jones et al., 2003). Jones et al. (2003) found that forecasters based flash flood threat assessments on variables including reflectivity, rainfall rates, rainfall totals, storm motion speed, hail contamination, and storm spotter observations. However, in the thirteen years since their study appeared, advances have been made in flash flood prediction and modeling, introducing forecasters to previously-unfamiliar ways of predicting this phenomenon. Although we do not debate that mental simulation and pattern recognition are still effective situation assessment mechanisms, we hypothesize that SA requirements may have adapted with the increasing availability of hydrologic-based flash flood forecasting decision support.

The Research Questions

Forecasters make sense of environmental situations by using computer-based decision support tools (also referred to as forecast guidance products) to compare between data sets continuously (Daipha, 2010; Kirschenbaum, 2004). Qualitative studies have established that forecasters' SA requirements vary throughout the forecast decision making timeline (Jones et al., 2003; Morss & Ralph, 2007). However, the practical implications regarding information use during flash flood forecasting are less understood; the same is true with respect to how SA requirements are affected by situational uncertainty. In the present study, we build upon Jones et al. (2003) and Lipshitz and Strauss (1997), using a mixed methods framework to assess SA requirements and cognitive processes in light of recent technological forecasting decision support tools. The first aim of this research was to explore the role of uncertainty in the weather forecasting decision making process (RQ2).

> *RQ2: How do forecasters build and maintain situation awareness while working under the constraints imposed by uncertainty leading up to a flash flooding event?*

In addition to aforementioned research question, we also explored the forecast decision making process including several proposed means of communicating forecast uncertainty. We aimed to develop a better understanding of the relationship between situation assessment and decision making under uncertainty by evaluating changes to SA requirements over time (RQ3). In this study, we assumed that the use of a forecast guidance product corresponded to an individual information requirement. This assumption was based on observations of forecasters using certain decision aids to gain particular types of information, as well as the knowledge that forecast information seeking is goal-directed (Hoffman & Coffey, 2004). Furthermore, we evaluated SA requirements not only at different time scales, but also at different levels of environmental activity; we anticipated that the amount of meteorological phenomena occurring on any given day would impact a forecaster's use of decision aids.

> *RQ3: Which tools do forecasters use, in combination and individually, to build situation awareness? How do their SA requirements change at different points along the forecasting compound warning decision process and at different environmental activity levels?*

Hypotheses

In relation to the aforementioned research questions, we expected to identify differences in SA requirements over varying time and environmental activity scales. Morss and Ralph (2007) observed forecast guidance use changing over time, and likewise, we hypothesized that flash flood forecasters would use different decision support tools at different times throughout the forecast decision making timeline. In an extension of this, we also hypothesized that SA requirements, as represented by the amount of time spent and frequency of decision aid use, would differ based on the environmental activity level. Finally, we predicted that the sensemaking process would not only involve a large degree of comparison, but it would also be a function of interpersonal communication, trust in decision aids, and personal background experience. The focus group study was exploratory in nature, so no testable hypotheses were developed. We expected that focus group discussions would produce insight into the comparative sensemaking processes forecasters use to assess an uncertain situation.

The Hazardous Weather Testbed Hydrology Experiment 2014

For decades, the testbed research framework has provided insight into forecaster decision making as well as forecast guidance efficacy (Clark et al., 2011; Heideman, Stewart, Moninger, & Reagan-Cirincione, 1993; Murphy & Daan, 1984). The method traditionally brings participants together to spend periods of time using forecast decision aids in mock-operational settings. A testbed's purpose is often to test new technological developments (Barthold et al., 2015; Clark et al., 2011), but some have the additional goal of addressing the role of technology and design in the decision making process (Heinselman, LaDue, & Lazrus, 2012; Karstens et al., 2015; Murphy & Daan, 1984). Testbed research has produced knowledge that encourages a research to operations (R2O) framework for technology development (Clark et al., 2011).

Motivated by its own impending transition from research to operations, a test of the FLASH system was conducted in July 2014. The Hazardous Weather Testbed Hydrology Experiment (HWT-Hydro) sought to evaluate a suite of hydrologic flash flood forecasting models while gathering knowledge about forecaster decision making processes. Using the suite of 30+ products, collectively known as MRMS-FLASH tools (Multi-Radar/Multi-Sensor, and Flooded Locations and Simulated Hydrographs, respectively), forecasters issued experimental watch and warning polygons throughout each of the four weeks during the experiment. In addition, HWT-Hydro occurred in coordination another experiment hosted by the Weather Prediction Center, the second Flash Flooding and Intense Rainfall Experiment (FFaIR; Barthold et al. (2015)).

HWT-Hydro occurred in four weekly cycles. Forecasters participated in weeklong shifts, and in each shift, a unique set of participants took part in the study. Upon arrival, participants received training on the use of the AWIPS-II weather forecasting display platform and the MRMS-FLASH tools, as well as an explanation of the general purpose of the experiment and research methods used throughout the week. The majority of the week was spent in real-time experimental forecasting operations. Each participant worked at an individual workstation, but usually partnered with a participant at a workstation near them in order to forecast over a shared geographic region. Due to the nature of the evaluation, participants were encouraged to rely primarily on the experimental tools, but they were allowed to consult external guidance tools online if the tools were not available on the testbed workstations. During the experimental operations, forecasters issued experimental watches and warnings across the continental United States. Participants in the WPC's FFaIR experiment provided a weather briefing to the HWT-Hydro forecasters in the form of a webinar at the beginning of each day.

Evaluation was addressed in a two-fold approach: (1) tool performance and forecast adequacy as well as (2) aspects of the forecaster decision making process. Tools and forecasts were evaluated in a subjective manner. Each day, participants completed a survey in which they evaluated flash flood events from the prior day's forecasts. The survey assessed how well the experimental tools predicted the actual threat, as represented by flash flood reports and other observations. The survey also analyzed experimental watch and warning spatial coverage, accuracy, and lead time in comparison to operational watches and warnings.

Participants also took part in a human factors-based, mixed methods analysis of warning decision making behavior. During forecasting operations, participants used desktop recording software to audio- and video-record their forecasting activities; the recordings were used for a time-based analysis of tool usage during the watch/warning issuance timeline. At the end of each week, participants took part in a focus group in which they gave feedback on the tools, discussed challenges in flash flood forecasting, and provided information about how experimental uncertainty attributes allowed them to communicate threat levels in their forecasts.

Study I: Quantitative Analysis Of Information Seeking Behavior

Method

Participants. Fifteen professional forecasters employed by the National Weather Service (NWS) participated in the 2014 Hazardous Weather Testbed Hydrology (HWT-Hydro) experiment. Out of the fifteen participants, eleven were male and four were female. While this is not as gender-balanced as might be desired, it may reflect the larger weather forecasting community, a field not known for its gender diversity (Daipha, 2010). Forecasters were from locations around the United States and worked for either a Weather Forecast Office (WFO; n=13) or a River Forecast Center (RFC; n=2).

Participants were selected one of two ways. In the first case, several participants were selected based on recommendations by supervisors within their home office.

However, most participants were selected through an application process and were selected based on background knowledge and interest in hydrology. As part of the application, potential participants wrote a short essay explaining their motivation for wishing to participate and their relevant qualifications. From those applications, the majority of the testbed participant pool was sampled based on their statements of interest and qualifications, such as professional role and education. Although not always listed in the essays, nine of the fifteen participants listed their professional role. Six participants were professional hydrologists, three held positions with the title of meteorologist, and the remaining six participants did not list their current job title.

Equipment, Materials, and Environment. The study took place in a controlled-access room located in the National Weather Center in Norman, Oklahoma. The testbed environment consisted of multiple dual-monitor computer workstations that were set up within the room. Each computer ran on a LINUX operating system and contained the Advanced Weather Interactive Processing System II (AWIPS-II) weather forecasting software in the Computer-Aided Visualization Environment (CAVE). Each computer also had the desktop recording software, RecordMyDesktop.gtk, installed upon it to facilitate data collection. Within AWIPS-II, forecasters had access to many forecast guidance products that served as decision aids. Though one of the testbed's purposes was to evaluate the performance of experimental guidance products, forecasters also had access to a nearly full set of operationally available products. A list of experimental products available in AWIPS-II during the HWT-Hydro testbed is shown in Table 8, and a full list of experimental and operational products used in the testbed can be found in Appendix B.

Table 8. Experimental (in development) flash flood decision support products

Decision Support Family	Decision Tool
Experimental Models	CREST Maximum Return Period HRRR-Forced CREST CREST Soil Moisture CREST Streamflow SAC-SMA Soil Moisture SAC-SMA Streamflow
Precipitable Water (PW)	Precipitable Water Analysis (RAOBs) Precipitable Water Percentile (RAOBs) Precipitable Water Analysis (RAP) Precipitable Water Percentile (RAP)
Quantitative Precipitation Estimate (QPE) & Quantitative Precipitation Forecast (QPF)	MRMS QPE MRMS QPF
Flash Flood Guidance Ratio (FFG)	QPE to Flash Flood Guidance Ratio QPF to Flash Flood Guidance Ratio
Average Recurrence Interval (ARI)	Precipitation Return Period (QPE) Precipitation Return Period (QPF)

Procedure. At the beginning of each day during the experiment, forecasters began by reviewing the weather conditions to establish situation awareness. Forecasters first received instructions to identify regions of the contiguous United States that may have had a threat level equal to a watch; severe weather watches are generally associated with long-term time frames, whereas severe weather warnings are short-term predictions that refer to impending threats. After issuing any watches, forecasters were then instructed to narrow their focus within a prescribed region to issue warnings, if necessary. For four to five hours a day, Monday through Thursday, forecasters used a set of operationally available forecasting tools along with the set of experimental tools to guide their decisions (in-development and therefore not available in operational NWS offices). Forecasters accessed operational tools in a number of ways, ranging from within the AWIPS-II platform to internet browsers. During all forecasting activities, participants' actions were recorded using the recordMyDesktop screen recording software. The software recorded forecaster interactions with the decision aids throughout the forecasting timeline; an example of a screen capture is shown in Figure 11.

The desktop recording software captured data related to decision aid use during situation assessment and decision making. Recordings totaled 186 hours and 36 minutes, but not all videos were of sufficient visual quality to distinguish participant interactions. Recording quality was poor at times, so blurred or choppy recordings were removed from the sampling population. Samples were taken from within thirty minutes to an hour prior to a participant issuing a watch or a warning. Sampling intervals within each recording varied on a case-by-case basis; this was because the sample start points were chosen to occur either at the beginning of a recording (only for those sampled

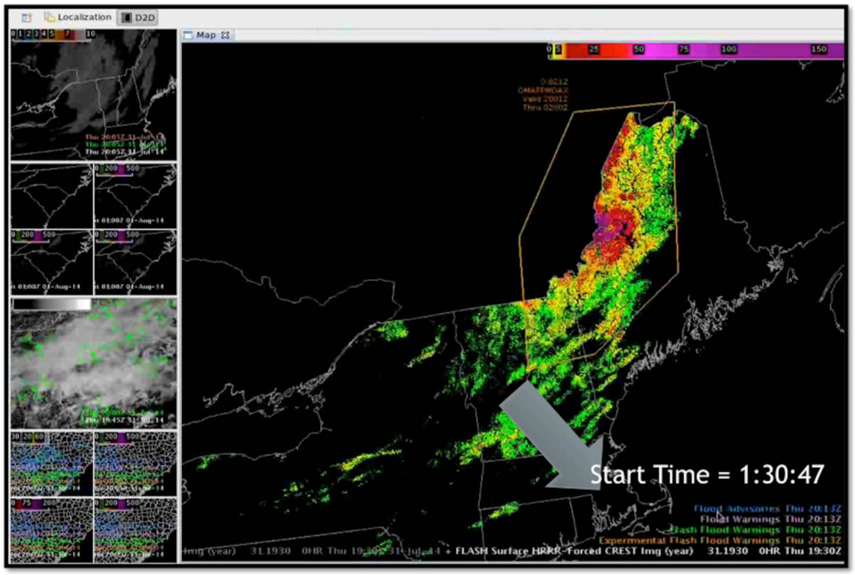
Localization
D2D
Map
25
50
75
100
150
Start Time = 1:30:47
Flood Advisories Thu 20:13Z
Flood Warnings Thu 20:13Z
Flash Flood Warnings Thu 20:13Z
Experimental Flash Flood Warnings Thu 20:13Z
FLASH Surface HRRR-Forced CREST Img (year) 31.1930 0HR Thu 19:30Z

Figure 11. Example of an AWIPS-II workstation display with multiple decision aiding visualizations present during the watch phase) or at a breakpoint in the decision process. Breakpoints were identified as either (1) the start of a new string of interactions following a prolonged break, or (2) the start of a new watch or warning product issuance following a prior product issuance. From the recordings that were of a sufficient quality, the sample consisted of 12 hours, 17 minutes, and 31 seconds (7% of the total recording duration). Separate analyses were conducted to assess SA requirements over time (watch issuance phase versus warning issuance phase) and environmental activity level.

Results and Analysis

Environmental activity level was based on flash flood, flood, and heavy rain local storm reports (LSRs) published in the Storm Events Database (National Climatic Data Center, 2014). Due to the situated nature of weather forecasting, using three types of weather events to represent environmental activity level was determined to be more appropriate than only looking at the number of flash flood reports. While the participants were only tasked with issuing forecast products for flash floods, floods and heavy rain can occur concurrently with flash flooding. Thus, a portion of the forecasting task would have involved detecting meteorological and hydrological patterns associated with flash flooding amidst heavy rainfall events. From the Storm Events Database, the number of water-related LSRs was calculated for each day in the experimental period. Percentiles were calculated from the LSRs, allowing for a three-level environmental activity scale (see Figure 12). Based on the July 2014 LSRs, the boundary between low and moderate activity was assigned to the 40th percentile, while the boundary between moderate and high activity was assigned to the 90th percentile.

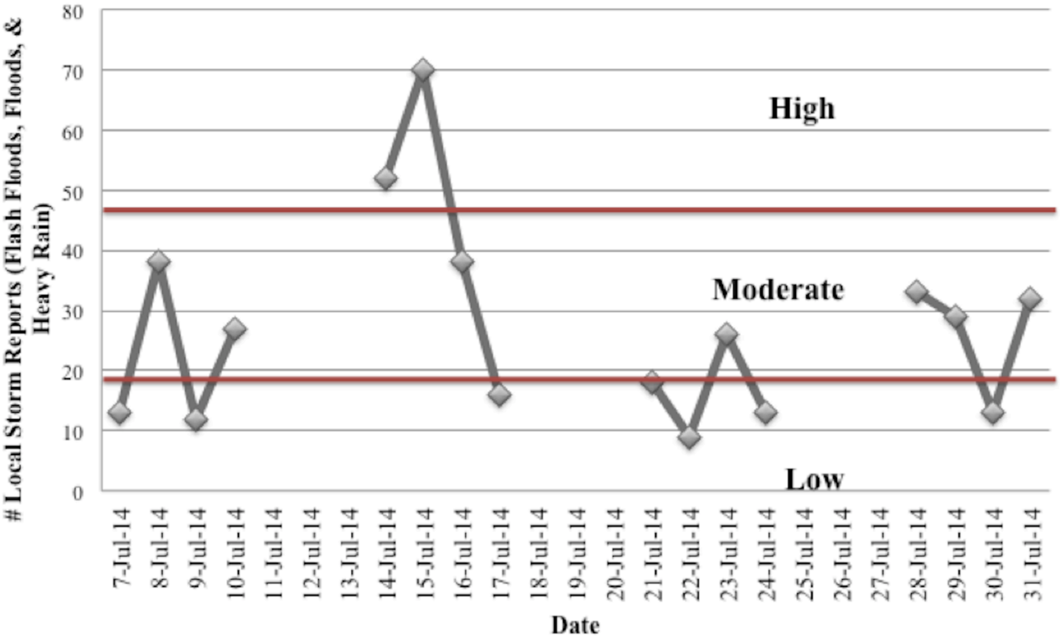

Figure 12. Number of Local Storm Reports (LSRs) for floods, flash floods, and heavy rain across the continental United States during the forecast period in July 2014. Gaps in the data show the days that experimental forecasting did not occur.

It is important to note that the scale measures environmental activity relative to the forecasts only in July 2014. Meteorological activity fluctuates in frequency and severity throughout the year and by geographic location, so what would be considered a "high activity day" in January is not the same as a "high activity day" in August. Likewise, what forecasters in Houston, Texas would view as a "busy" day is likely not the same as what a forecaster in Akron, Ohio would consider one to be. Due to the nature of meteorological phenomena, environmental activity level was not a controllable factor, as one can see in Figure 12. However, this was addressed through balanced sampling methods when selecting segments from the screen recordings.

The recordings produced several conclusions related to SA requirements during the watch and warning issuance timeframes through an analysis of the time and frequency of forecast guidance usage. During the analysis, the video samples were transcribed, which involved watching the samples and recording the start time and end time associated with each tool's use. For example, in Figure 11, the participant had just chosen to display the FLASH Surface HRRR-Forced CREST model. In the transcription process, the product's name and the timestamp at which it was first displayed (t = 1:30:47) was transcribed. Then, when the participant chose to remove the FLASH Surface HRRR-Forced CREST model from the central panel, the timestamp at which the removal occurred was transcribed. Should the same product have been displayed a second time later during the forecasting process, a new entry with start and end time would have been recorded. An example of one of the video transcripts is included in Appendix C.

Watch Issuance. In the videos sampled during the watch issuance forecasting activities, participants used a mean of 20.1 forecast guidance products (σ = 9.4). As shown in Table 9, ten samples were taken from the population of recordings. These samples represent 6 hours, 5 minutes, and 43 seconds worth of data.

In order to assess "big-picture" SA requirements, a measure of cumulative time was determined for each guidance type; for a full list of individual products and their respective guidance type. While analyzing forecaster behavior with regard to tool usage, several things become apparent. Figure 14 represents the cumulative time the participants spent using groups of forecast guidance products. For example, the "radar" category is the sum of the times spent by all participants using both available radar tools (the FLASH Surface MRMS Seamless Hybrid-Scan Reflectivity as well as the FLASH Surface MRMS Quality-Controlled Composite Reflectivity). As a result, this leads to

Table 9. Number and duration of sampled desktop recordings for flash flood watch forecasts

	n	Duration (min.)
Low	3	97.53
Moderate	3	150.78
High	4	117.40
Total	10	365.72

the measure reflecting the total time that all products within each guidance family were on the screen in any combination.

In Figure 13, one can see the cumulative time that each guidance product was visible on a participant's screen. Although the sample contained just over six hours worth of forecasting, the combined screen time of products in the "FFA/FFW/LSRs" category was more than eight hours; this indicates that the products in the "FFA/FFW/LSRs" category were used for extended, overlapping periods of time, and so when taken in combination, the total time exceeds the sample time. Looking at Figure 14, the total screen time given to products in the "FFA/FFW/LSRs" category grossly outweighed any of the other guidance product categories. However, it is important to note that the "FFA/FFW/LSRs" products were not predictive tools; they instead provided general SA in terms of existing flash flood warnings, both operationally and as issued by other study participants. These tools were often overlaid on top of prediction models and radar imagery.

Apart from the "FFA/FFW/LSRs" products, radar imagery, hydrologic data (in the experimental models), and flash flood guidance (FFG)-based models ranked in the top three most-used guidance tools. In terms of SA requirements for watch-phase decision making, the data show that participants placed heavy focus on these decision

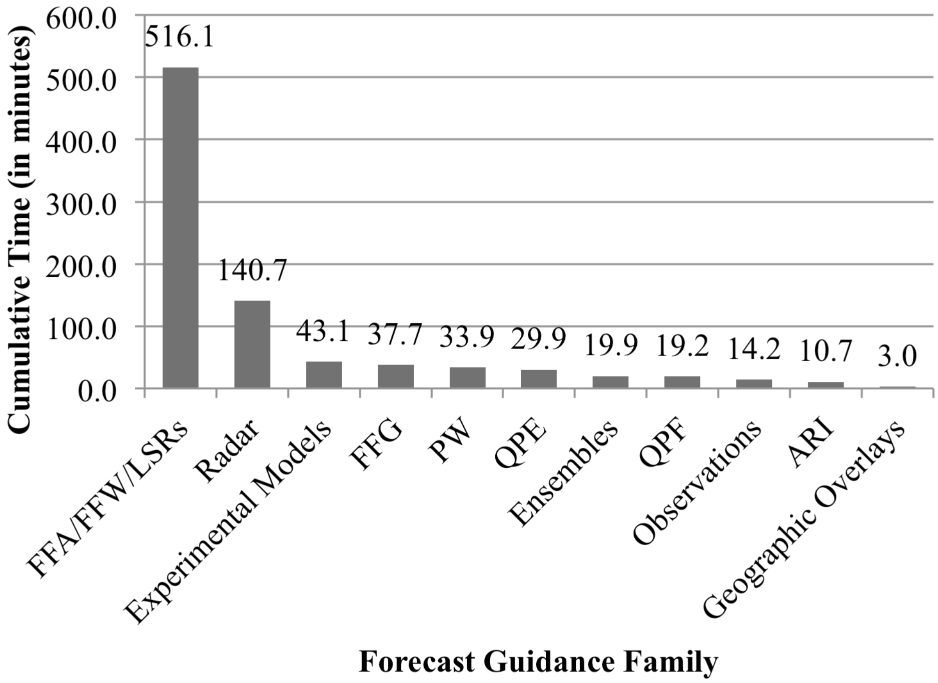

Figure 13. Cumulative time spent using forecast guidance products available in AWIPS-II (for issuing watches)

aids. The radar products and FFG-based models may have received a larger amount of screen time due to their prominence in operational forecasting; traditional methods of flash flood forecasting encourage forecasters to assess radar products for heavy rainfall signatures, and FFG-based decision aids have been used across the NWS for several decades. Thus, it is not entirely surprising that study participants relied heavily on familiar tools that they trusted. Nevertheless, the experimental hydrologic models—forecast guidance that participants did not have prior experience using—received a relatively large amount of screen time compared to the other guidance categories. While it is possible that this measure was affected by HWT goals (to evaluate experimental guidance products), we believe that the effect was not significant. Accordingly, other experimental products, such as the average recurrence interval models (ARI) or quantitative precipitation forecasts (QPF), did not receive the same amount of attention. Instead, we conclude that the experimental hydrologic models provided new and useful information in the watch decision making process.

The data also confirmed that during watch-phase decision making, SA requirements differed as environmental activity level changed. Figure 14 represents the proportion of time participants spent using forecast guidance products relative to environmental activity level. Although the *x*-axis is ordered in line with the temporal order in Figure 13, time has been normalized in Figure 14 in order to reflect differences between environmental activity levels. As a result, several variances appear in forecast guidance usage. During moderately active days, participants spent more time evaluating experimental models, FFG ratio guidance products, and ensemble models than they did during high or low activity days. With respect to usage of experimental models, it is hypothesized that this is due to a more balanced workload during

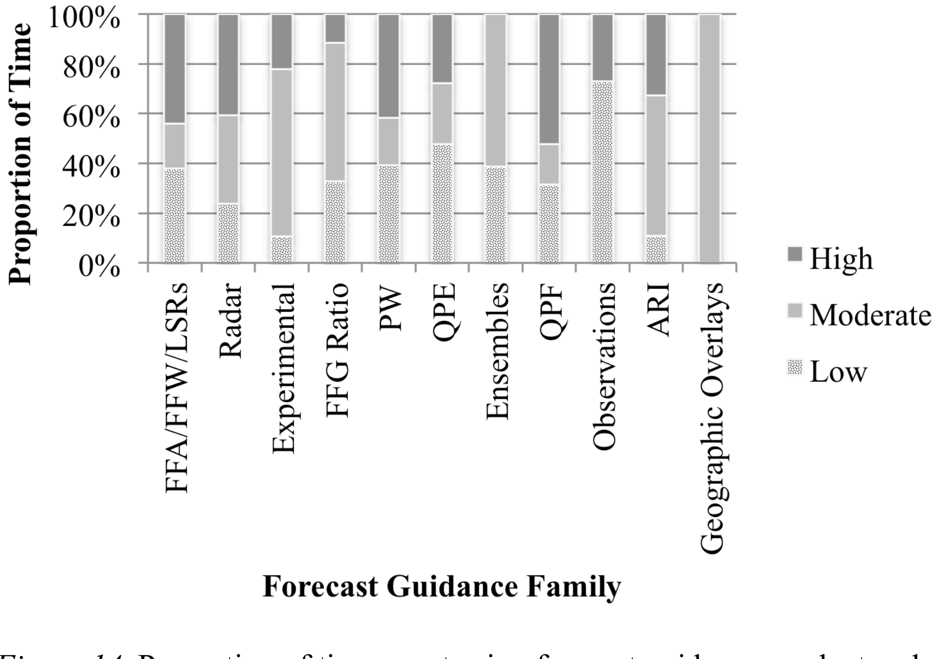

Figure 14. Proportion of time spent using forecast guidance products relative to environmental activity level

moderately busy days than during either high or low activity days. During highly active days, participants had many claims upon their attention, and so they may have relied more on familiar guidance products in order to reduce workload and stress. However, it is less clear why experimental models received less usage during low activity days. It is possible that the lower screen time was due to lack of modeling; by definition, little environmental activity occurred on “low activity” days, and as such, the hydrologic models would have been less likely to predict events. This may explain the spike in usage for observations on low-activity days; although observations did not receive as much screen time as the experimental models, participants may have found observation maps more informative than blank-map models. Interestingly, we see that during the sample, participants only used customized geographic overlays (such as adding river maps to the background display) during the moderate activity level days. However, in Figure 13 one can see that the geographic overlays had the lowest cumulative time, and as the data came from only one participant in the sample, it is hypothesized that using geographic overlays is a personal choice that can be attributed to individual differences.

Conversely, usage of several products declined during moderately active days with respect to their application during high- and low-activity days. Precipitable water (PW) products, quantitative precipitation estimate (QPE), and quantitative precipitation forecast (QPF) products received much less screen time during average days than during the extremes. However, the nature of these products and the watch forecasting process may shed some light on these differences.

Watches and warnings differ in timeframe. A watch refers to a general threat that may or may not occur usually 6 to 24+ hours past the point of issuance, whereas a warning refers to a specific threat that is likely to occur within 6 hours or less of the point of issuance. PW is a measure of the mass of water held within a column of the atmosphere (American Meteorological Society, 2015), and it is based on atmospheric soundings (rawinsonde observations; RAOBs) and the Rapid Refresh (RAP) numerical forecast model. It is updated on an hourly basis, and as such, gives short-term atmospheric information to users. During the watch issuance timeframe, forecasters are trained to assess whether atmospheric and environmental conditions are consistent with specific threats. As such, SA requirements for a watch will include information about such conditions; in terms of decision aids, this translates into guidance products that provide users with data about ground-based observations and atmospheric conditions, such as the PW-based decision aids.

In addition to interactions with forecast guidance products, the recordings revealed that participants engaged in several other types of activity during the forecast process. A large portion of time was spent transitioning between display screens, setting up new layouts, and reviewing geographic locations on blank maps. The total time spent doing these types of activities was 57 minutes. Participants were able to view operationally issued watches and warnings in the AWIPS-II display; this capability not only allowed participants to display the watch and warning polygons alongside the other visualizations, but it also gave them access to official operational text products. The text products contained professional discussions regarding justifications for the forecast, and it is possible that this information was beneficial to testbed participants in that it contributed to their situation assessment process. However

Table 10. Number and duration of sampled desktop recordings for flash flood warning forecasts

	n	Duration (min.)
Low	2	116.52
Moderate	3	124.33
High	3	130.95
Total	8	371.80

in the sampled videos, the total time spent reading these discussions was just over two minutes (t = 2.47 minutes).

Warning Issuance. As with the watch phase analysis, approximately six hours of screen recordings were sampled across 8 videos within the total data set, as shown in Table 10. Likewise, samples were selected from the subgroups of environmental activity level. Participants consulted slightly more unique forecast guidance products in minutes leading up to issuing a warning than prior to issuing a watch (μ = 25.8 guidance products per day, σ = 6.3). As hypothesized, SA requirements in the warning decision making process differed from those required for watch decision making. Figure 15 presents the temporal analysis results from the recordings sampled.

Similar to the watch decision making process, the "FFA/FFW/LSRs" overlays were used frequently in combination with each other and with other products, as indicated by the cumulative time metric exceeding the total duration of the sample. The large amount of screen time given to the "FFA/FFW/LSRs" category in both watches and warnings suggests that the overlays assisted participants in maintaining SA in terms of existing operational and experimental forecast issuances. Apart from the forecast overlays, FFG-based guidance products, radar imagery, and QPE-based products received the greatest amount of screen time. FFG products' high amount of screen time

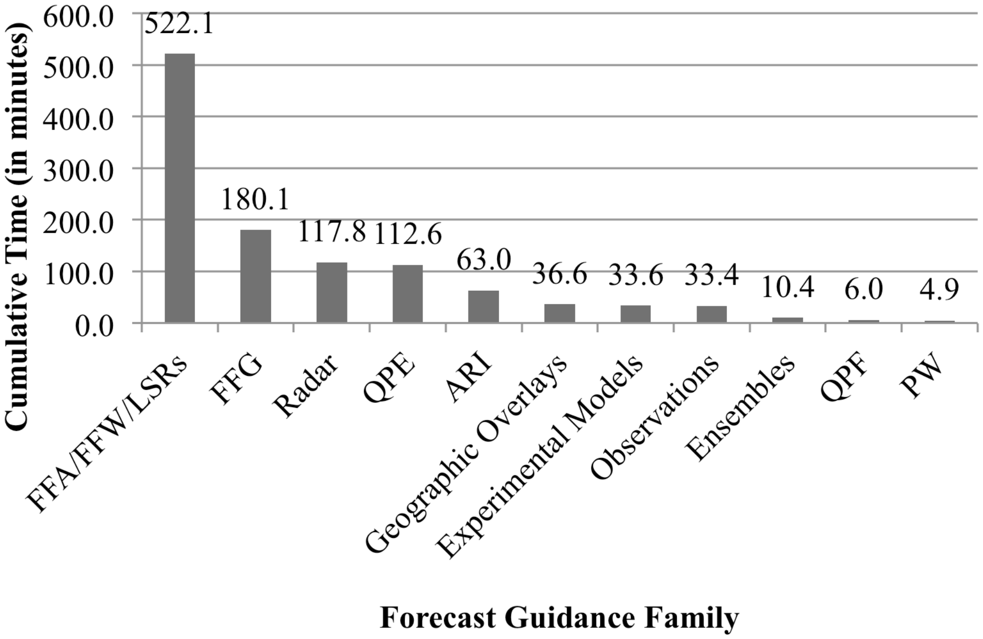

Figure 15. Cumulative time spent using forecast guidance products available in AWIPS-II (for issuing warnings)

is not unexpected; although several of the FFG-based tools were experimental (the FFG ratio guidance products), participants were experienced users of operational FFG decision aids. Thus, familiarity with traditional FFG decision aids may have led participants to put greater trust in the testbed's experimental FFG products. When used in combination with unfamiliar decision aids, participants may have used the FFG-based products to manage individual uncertainty with regard to the bias and prediction outputs of the other experimental models.

As shown in Figure 16, the relationship between decision aid usage during the watch and warning phases did differ. The forecast guidance categories are ordered from greatest to least cumulative screen time, as was previously shown in Figure 15. Several types of guidance were used somewhat equally between watch and warning analysis, namely the radar imagery, and the experimental hydrologic models. However, there

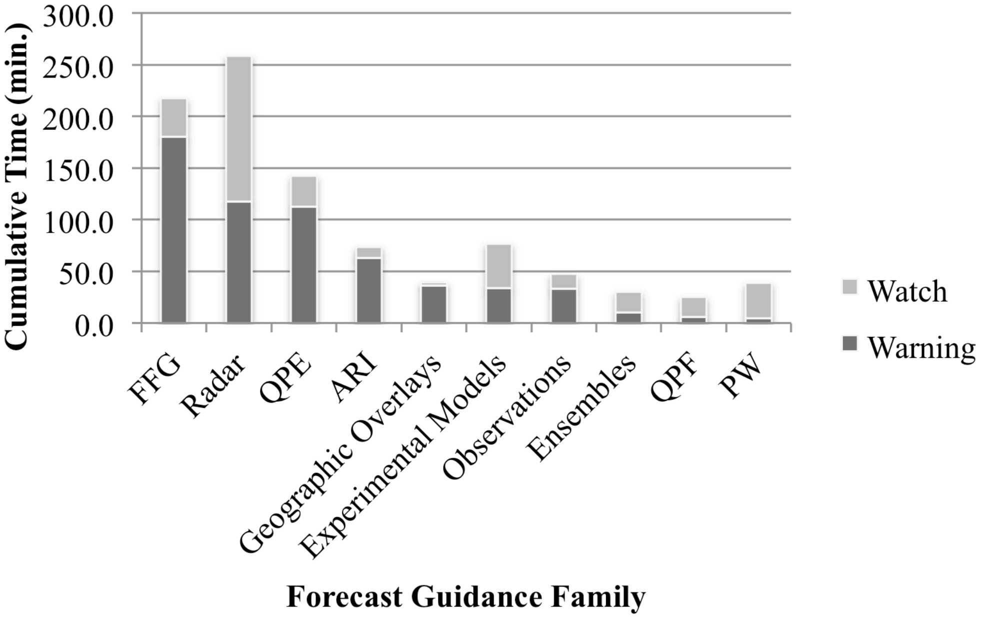

Figure 16. Cumulative time spent using forecast guidance products during watch and warning phases, with "FFA/FFW/LSRs" category removed to reveal differences in the lower end of the spectrum

were some observable differences in usage between the two phases along the decision timeline. For some products, usage decreased as the decision lead-time decreased.

While QPF decision aids were one of the less-viewed categories during the watch timeframe, their perceived utility further decreased in the warning timeframe. Whether this was due to a lack of adequate training on QPF interpretation or a different reason would require further investigation. In addition, precipitable water (PW) guidance products decreased in the warning phase, becoming the category with the least screen time prior to warning issuance. PW is modeled over large spatial scales, and outputs change slowly over time; these qualities give PW-based decision aids greater utility during the watch timeframe, where watch forecasts are also issued over larger temporal and spatial scales than warning forecasts. This is also a likely explanation for the proportional difference in the screen time given to ensembles; ensemble models are visualized as “spaghetti plots,” or mappings of overlapping lines that represent atmospheric pressure levels. Like PW, pressure is an atmospheric condition that informs decisions about big-picture risk for meteorological phenomena, and less about discrete environmental threats.

Conversely, guidance products based on ARI and QPE contributed to warning decisions, which is indicated by a greater proportion of screen time during the warning phase than in the watch phase. While they received more screen time in the warning phase, however, they were still some of the less-used products overall. It is logical to assume that QPE and ARI information satisfied requirements for building SA. ARI products provide users with information about frequency of floods with a specified size. The proportional increase in ARI guidance usage for warning decisions indicates that knowledge about risk level is a component of good SA for flash flood warning forecasts. While this would likely be useful for forecasters to know when predicting flash floods in the long term, the modeling involved in predicting ARI limits the ability to provide such information to forecasters; rainfall estimates are one of the inputs into the ARI model, so the ARI products are only able to provide measures of risk after rain has already begun. During the watch phase, the decision making timeframe occurs too far in advance of rainfall to create outputs in the ARI guidance; this also is the case for QPE outputs, and likely accounts for their low usage during the watch timeframe.

While QPE received more screen time during the warning timeframe, the reverse was true for the quantitative precipitation forecast (QPF) guidance products. Indeed, as shown in Figure 16, the actual time that participants used QPE products was much greater than that spent with QPF products. However, proportionally, the screen time given to QPF products indicates that they had more utility as a decision aid in the watch timeframe. This is supported by the guidance product's design; within HWT-Hydro, the QPF products provided estimates of precipitation amounts based on QPE, and were projected as 15-hour outlooks. This long-term outlook made QPF a better fit for guiding risk assessment in the long-term.

In line with the watch phase analysis, forecast guidance usage in the warning phase also changed with environmental activity. In Figure 17, one can see that FFG-based decision aids received more screen time during low activity days than in either the moderately or highly active days. Likewise, while geographic overlays and ensemble products did not receive a majority of the screen time, when they were used, they were most often used on low activity days. This may be due to effects of participant boredom. During less active forecasting periods, participants would have

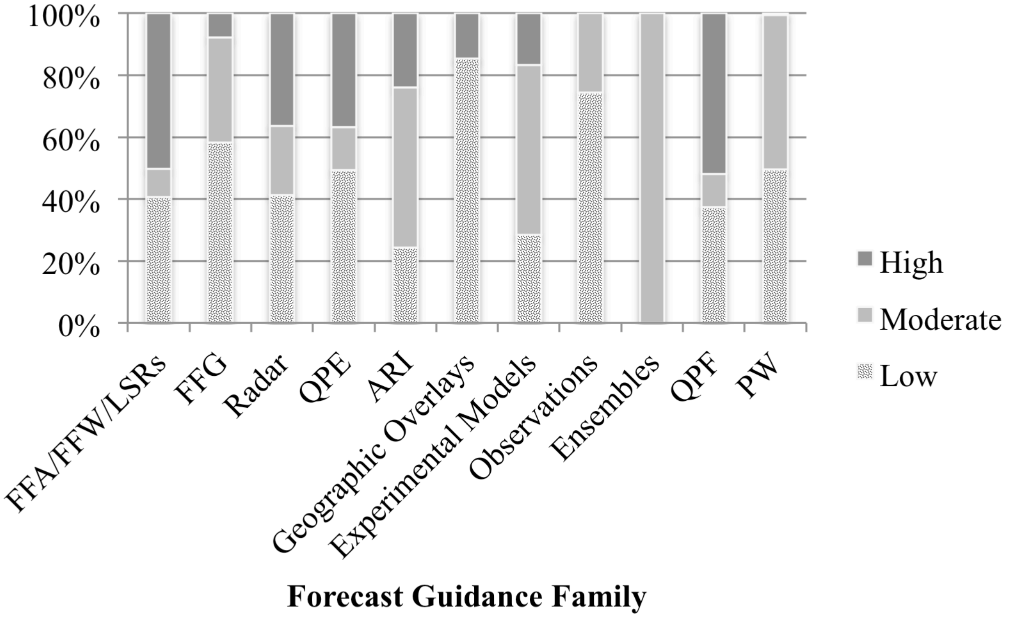

Figure 17. Proportion of time spent using forecast guidance products during warning issuance relative to environmental activity level

had fewer threats to monitor, and would in turn have more time become familiar with local terrain and geographic features, such as burn scars, that would increase the risk of flash flooding within their forecast domain. Interestingly, the experimental hydrologic models received the most screen time on highly active forecasting days. This may suggest that hydrologic data was an important part of the situation assessment process; it is possible that they provided a way to distinguish significant threats from noise.

Time-Frequency Analysis of Experimental Decision Support. In addition to a temporal analysis of forecast guidance usage, frequency of use was recorded and analyzed. Frequency was measured by counting the number of times a participant added a specific decision aid to his or her AWIPS-II display. An example of product frequency can be found in the data transcript (Appendix C). In order to assess the relationship among the experimental decision aids, situation assessment, and the forecasting timeline, a time-frequency analysis was conducted; the results are shown in Figure 18, where cumulative time (y-axis) is presented as a ratio of screen time to total sample time, and frequency (x-axis) is presented as the percentage of product-specific interactions to total interactions.

In line with the temporal analysis, the FFG-based guidance products not only received the greatest screen time, but they were also the most frequently selected. It is important to note that a high frequency does not guarantee a high rate of use; throughout the testbed, participants were regularly observed engaging in rapid comparison activities. These rapid comparisons involved switching between views, toggling between one product and another (or several). In the frequency analysis, a high frequency value more often than not indicates that the product in question was

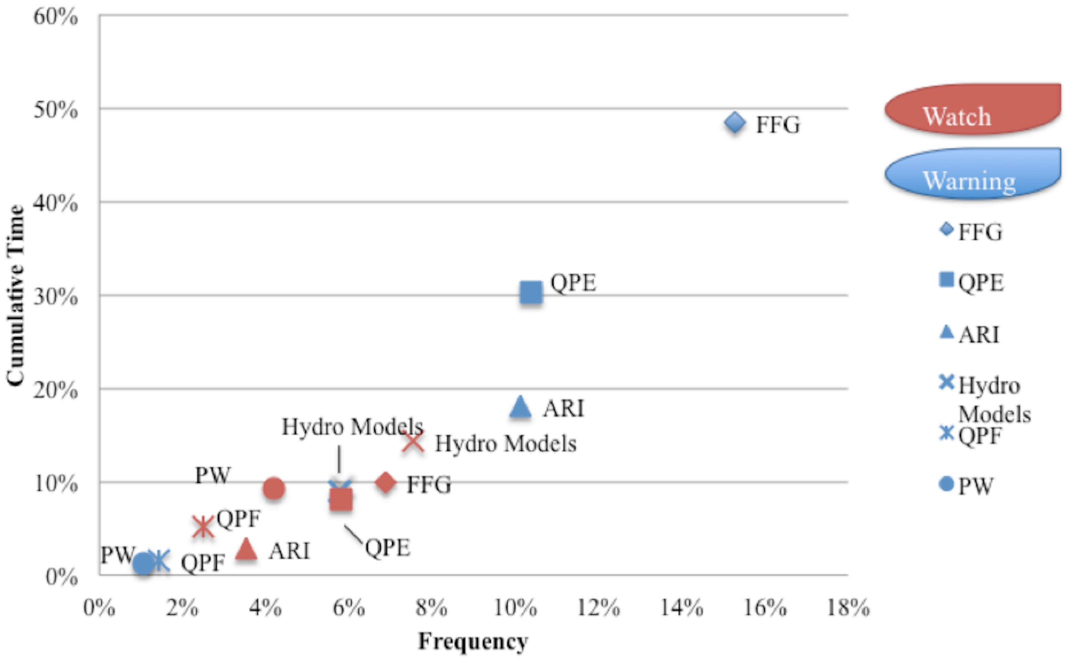

Figure 18. Time-frequency analysis of experimental guidance product usage during watch and warning issuance

regularly used as a comparative tool alongside others. Therefore, one could conclude that while PW and FFG-based products had roughly equivalent screen time during the watch phase, the FFG-based products were used as comparative decision aids more often than PW products.

In the warning phase, the QPE- and ARI-based products had similar frequencies of use, but received different amounts of screen time. This may suggest a difference in usability. QPE products were selected just as frequently as ARI products, but the greater screen time indicates that participants either found certain products to be more useful or that they required more time to incorporate the information into their situation assessment and sensemaking process.

Discussion of Study I

The time-frequency analysis indicates that guidance usage was dependent upon the forecast lead-time and environmental activity level, confirming the original hypotheses. During the long-term watch issuance timeframe, we see several differences between activity level and reliance. In a review of the amount of time spent viewing each decision aid, we see that the forecasters spent the most time looking at non-experimental tools. However, if one disregards the prevalent use of FFA/FFW/LSR overlays and the radar imagery, the data suggest that hydrologic models and experimental products were used more frequently than operational products. Indeed, on moderate activity days, forecasters spent nearly an hour more with non-experimental tools visualized in their displays.

A different behavioral pattern emerged during the warning timeframe. Out of the experimental decision aids, participants relied on tools they were more familiar with, such as FFG ratio maps, QPE products, and ARI outputs. Likewise, in terms of operational decision aids, participants spent more time viewing observation reports during the short-term warning phase than in the watch phase, which is consistent with the view that flash flood observations contribute to situation awareness. On the days with moderate environmental activity, participants spent about half as much time using operational tools as they did during either high or low environmental activity days. This is not surprising from a tool development perspective, as the experimental tools were designed to aid in the short term forecasting stage.

The findings confirm those of Morss and Ralph (2007), who also observed that forecasters accessed different guidance tools throughout the forecast timeline. In addition, the present study’s findings strongly suggest that guidance usage also differs depending on day-to-day activity level. From a decision making perspective, guidance usage over time is a useful, albeit nontraditional, identifier of SA requirements. Differences in usage frequency and screen time reflect changing user needs with regard to information sources necessary to predict flash flooding. However, a high amount of screen time does not necessarily translate to more importance as an SA requirement; a prime example of this is the relatively low amount of time given to geographic overlays compared to other decision aids. Morss and Ralph (2007) discussed the importance of local knowledge when producing weather forecasts, yet in the present study, the geographic overlays received the most use during low activity forecasting days. While still an SA requirement for busier forecasting periods, we hypothesize that participants prepared during low activity days, consulting geographic and topographic information sources to build their local knowledge in advance of the highly active periods. In this way, forecasters shifted some of the workload out of the busier shifts and used downtime to build up memory stores for when they were needed.

One must consider several things that limit our conclusions and generalize the data. In spite of or perhaps because of the training given at the beginning of each week during HWT-Hydro, some of the forecast guidance tool usage patterns may have been biased. On the one hand, it was hypothesized that experimental model usage may have been higher than it would have otherwise been due to the participants’ awareness of the testbed’s goals (to evaluate experimental products). However, after the temporal evaluation, it is clear that while experimental decision aids were used widely, not all were given equal screen time; indeed, traditional sources, such as radar imagery, were still used more widely than any of the hydrologic experimental models, ARI guidance, or QPF guidance. On the other hand, though, a reverse explanation may also be possible. Although participants received training about each experimental product prior to the first forecasting session, it is possible that participants were still uncertain regarding appropriate usage of the experimental tools. If this were the case, potentially inadequate training may have led to the lesser screen time given to products like QPF- and PW-based decision aids.

Study II: Forecasters' Management of Uncertainty and the Forecast Decision Making Process in HWT-Hydro 2014

The qualitative component of the mixed methods study sought to identify processes involved with maintaining situation awareness under uncertainty during flash flood forecasting (RQ2). A focus group was used to collect open-ended responses regarding participants' uncertainty management techniques, forecast guidance usage, and feedback on a set of experimental risk communication methods for flash flood forecasting. The following section presents a discussion of the method, followed by findings from the focus group discussions.

Method

Participants. Fifteen participants took part in the focus groups during HWT-Hydro 2014; these were the same participants that took part in Study I. Participants were all National Weather Service forecasters, with either primary job roles in hydrologic or meteorological forecasting. They primarily worked at Weather Forecast Offices (n = 13), though a few were based out of River Forecast Centers (n = 2).

Forecasters were selected from offices and centers around the continental United States, so a wide variety of geographic regions were represented in the sample.

Each focus group consisted of three to four individuals. Although the recommended minimum number of participants per group is six due to concerns of reduced conversation (Caplan, 1990), discussions lasted about one hour and responses flowed naturally. In terms of background, participants' forecasting backgrounds qualified them to discuss the group themes related to coping with uncertainty. While the groups were homogenous in terms of forecasting qualifications, participants represented Weather Forecast Offices and River Forecast Centers across the continental United States. Differences between office cultures, regional forecasting policies, and responsibilities within the National Weather Service were all anticipated. These differences stimulated discussion and participant interaction during the group meetings.

Focus Group Design. Focus groups are often used in human factors research because of their ability to provide highly detailed, qualitative data about a central theme. In a focus group, individuals explore perspectives on an idea or product, guided through the discussion by a moderator. Moderators must have subject matter experience and the ability to facilitate discussions while minimizing experimenter bias (Caplan, 1990). The focus group methodology is particularly suited to capture information about participant beliefs and attitudes through analysis of individual responses and group interactions (Freeman, 2006). Philosophical perspectives differ in the importance of group homogeneity and the importance of interpersonal interactions during discussions. As discussed by Freeman (2006), the contextual constructionist perspective cautions against homogenous groups due to the assumption that group member similarities constrains discussions (Kitzinger (1994) as cited in Freeman (2006)). Conversely, the realist perspective recommends group homogeneity, in that within-group similarities allow for intra-group comparisons and in turn, promotes external validity (Kreuger and Casey (1994) as cited in Freeman (2006)). Both perspectives recognize the face validity of interaction analysis; interaction promotes discussion in the realist philosophy and is the source of meaning from the constructionist perspective (Belzile & Öberg, 2012; Freeman, 2006)

From a practice-oriented perspective, focus group design can impact the validity of the results. Specifically, personal attributes of the moderator can affect discussion outcomes in terms of how group members perceive and relate to the discussion leader (Belzile & Öberg, 2012). For ergonomics-related focus groups, Caplan (1990) recommends that moderators have a background relevant to the focus group's subject matter, facilitation experience, and neutrality about the topic to minimize bias.

In the present study, three different individuals moderated the focus groups: moderator M1 facilitated discussions with group 1 and 4, moderator M2 facilitated discussions with group 2, and moderator M3 facilitated discussions with group 3. All moderators were graduate students in their mid-twenties, but M1 was female while M2 and M3 were male. The majority of the group participants were male, so theoretically some gender bias to the responses may have existed; however, this was not observed and any potential effects were subtle.

Focus groups met at the end of each week during the testbed study. During the discussions, participants and the moderator sat around an oval conference table to encourage interaction between members. The focus group addressed a range of topics related to the general forecasting process as well as the participants' views on uncertainty, probability, and confidence in flash flood forecasting. The questions of interest to the present study were those that sought to elicit feedback on the role of the uncertainty attributes in communicating threat information to end users. During the group discussions, the questions were posed as:

- General forecasting background and experience (2 questions)
- Decision making under uncertainty (5 questions)
- Using impact characterizations in forecasting communications (5 questions)

A full list of focus group questions can be found in Appendix C. Questions were designed to elicit open-ended responses and follow-up questions or comments by participants were encouraged to stimulate discussion. In addition, questions were piloted with a test group of subject matter experts prior to the formal group meetings. The general forecasting background questions were designed to engage and introduce participants to the discussion topic. Nine of the questions related to decision making and the impact characterizations were designed to explore the central theme of coping with and communicating uncertainty. At the conclusion of each group, the final query acted as an exit question to capture anything that may have been missed in earlier discussions.

Thematic Analysis Protocol. Focus group discussions were audio-recorded and then transcribed. The transcripts were then analyzed using thematic analysis. Thematic analysis, a form of qualitative content analysis (QCA), is a flexible, systematic methodology used for capturing themes and patterns within a qualitative dataset (Schreier, 2012). Themes represent elements of the central organizing concept for the analysis, and are often identified as prevalent patterns of responses (Braun & Clarke, 2006). Themes are derived from categories of codes, defined by Saldana (2015) as "a short word or phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data." In practice, codes are words or statements that briefly paraphrase ideas, emotions, or behaviors expressed in the data. In thematic analysis and other QCA methods, analysis is an iterative process involving an initial coding stage, code refinement, refinement of codes into themes, and structural verification (Braun, Clarke, & Terry, 2014). The present study draws upon the methodological framework outlined by Braun and Clarke (2006), who present a highly detailed guide to conducting a thematic analysis. A summary of the steps involved in this type of research is presented in Table 11.

In thematic analysis, the researcher takes an active role in identifying and interpreting meaning in the information; as such, it is important to acknowledge the research epistemology prior to and during analysis as it impacts the types of conclusions that can be drawn (Braun & Clarke, 2006). In the present study, the author identified with the realist perspective and employed a theoretical and semantic analytical approach. In a semantic approach, the researcher summarizes themes, interpreting patterns and identifying relationships between themes, particularly in relation to existing theoretical frameworks relevant to the research question; in this approach, meaning is drawn strictly from the researcher's interpretation of participant responses (Braun & Clarke, 2006). The semantic approach contrasts to the alternative latent-level

Table 11. Steps for conducting a thematic analysis, summarized from Braun and Clarke (2006)

Phase	Activities
(1) Familiarization	Transcribe and review data (transcripts, media, etc.)
(2) Initial Coding	Identify codes and patterns in the data
(3) Search for Themes	Collapse codes into themes, create initial thematic hierarchy
(4) Review Themes	Test thematic hierarchy, review and refine themes to create thematic map
(5) Define Themes	Finalize inclusion rules for themes, select representative names for each theme
(6) Report	Illustrate thematic structure with examples from the data that relate to research question

thematic approach, in which the researcher attempts to interpret the underlying motivations for participants' use of particular semantics.

Complementing the semantic approach, the author also assumed a theoretical approach to coding. The theoretical approach, which is grounded in an existing theoretical framework, allows the researcher to explore the data through a predefined lens. This produces highly detailed findings related to the core research question (Braun & Clarke, 2006). The theoretical approach facilitated a top-down analysis of how forecasters cope with uncertainty, driven by the tactical framework discussed by Lipshitz and Strauss (1997) and by the macrocognitive model of forecast decision making presented by Trafton and Hoffman (2007).

Findings from the Thematic Analysis

In thematic analysis, themes are often associated with measures of prevalence, which can be assessed in terms of a theme's presence across the entire dataset, presence in individual sources within the dataset, or as a reference frequency measure, which captures the number of times a topic was mentioned in the course of the dataset (Braun

& Clarke, 2006). In the present study, the researcher analyzed the data both in terms of presence of themes by focus group and in terms of thematic frequency. The frequency measures, similar to those used in other QCA methods, allowed the researcher to draw conclusions related to the relative importance of themes. However, Braun and Clarke (2006) express the concern that frequency measures can be difficult to apply due to issues created by the size of units of analysis. For this reason, the analysis also discusses themes with a presence/absence variable between focus groups.

General Forecasting Background. Several questions probed participants about their individual experience with flash flood forecasting. These questions were designed to elicit information regarding how participants operationally use forecasting decision aids in the warning decision making process. Discussions not only revealed general situation assessment procedures, but they also reflected participant perceptions of the importance of the role of the forecaster in the weather domain. In one particular focus group, this theme of forecaster self-image emerged particularly strongly: participants agreed that as forecasters, their roles involved acting as a "weather authority" and as "communicators." These viewpoints demonstrate the importance of not only having meteorological knowledge and the ability to create a forecast, but also the value in being able to provide information to forecast end users. As one participant noted, the forecaster's role increasingly overlaps with the role of a decision support service.

Decision Making Under Uncertainty. After using the background experience questions to establish an environment conducive to holding open discussion, the conversation turned towards more specific aspects of the warning decision making process. Participants discussed their own methods for making sense of forecast information, and they were encouraged to give specific examples from the testbed and from their home offices. Analysis of the discussions produced a number of themes relating to situation assessment and action choice during the watch and warning decision making process.

Establishing the Big Picture. At the beginning of any forecasting shift, participants agreed that the decision space was often a "blank slate." The decision making process begins by attempting to *establish the big picture* in relation to the environmental and atmospheric states. In order to transition from a blank slate to understanding the big picture, forecasters *assess environmental parameters* between and within information sources. Preferred information sources differed from forecaster to forecaster, but several guidance products that were frequently mentioned for flash flood forecasting were radar imagery, precipitable water (PW) estimates, vapor imagery, and atmospheric soundings. Such sources provided a coarse level of detail, but they also allowed the forecaster to form a baseline for flash flood risk assessment.

Participants pointed out the importance of maintaining an awareness of the big picture throughout the entire forecast decision making process. Although participants did not use such wording, this baseline understanding may in fact relate to situation awareness. One participant spoke of the importance of maintaining awareness of the environmental baseline conditions, stating that in the testbed, he "always felt like [he] had enough time to pull back and look at the big picture." Using a cyclical process, this forecaster used his SA of baseline conditions to develop SA more focused on specific flash flood threats.

Focusing Attention. The second stage of forecast decision making involved source comparison and parameter assessment to *focus attention* on specific threats. Similar to the "big picture" stage, SA was built through assessment of environmental parameters. However, when discussing this stage, participants largely agreed that assessment occurred in two contexts. In the first context, assessment occurred within individual guidance product; forecasters sought specific guidance products and evaluated predictions in terms of thresholds (e.g. flash flooding may occur when the QPE exceeds a certain amount in a certain location), societal conditions (e.g. local infrastructure or vulnerable populations), and model bias estimates.

The second context was more comparative in nature. Participants discussed building SA by comparing between guidance products; it is possible that this process served to refine the forecasters' mental models. One participant stated that she "would look at the flash flood guidance and kind of switch between [that and rainfall return periods], but a lot of times they were both showing about the same story, and the rainfall return periods were better at providing an estimate of magnitude and scope." If the within-product assessment could be likened to Endsley's Level 1 SA (perception), between-product assessments could be seen as similar to Level 2 SA (comprehension).

While parameter assessment was one of the most frequently discussed elements of developing SA, group participants also acknowledged the role of *interpersonal communication* in the situation assessment process. Particularly during the testbed, participants relied heavily on the briefings from the associated testbed to direct their attention to certain regions across the country. Operationally, forecasters are accustomed to seeking advice from colleagues; during the focus groups, one participant gave the example that, “there’s always going to be different opinions… on model solutions, so there’s…. there’s a lot of discussion that happens to get to [an agreement].” In addition, negotiation and discussions between offices may occur in order to issue a forecast product over a large geographic region. Discussions revealed that such processes allowed forecasters to identify patterns over time and space, which in turn created the confidence needed to proceed to the action stage.

Action Selection. In the final decision making stage, forecasters activate the knowledge developed in the earlier stages as part of the action selection process. This stage involved a number of interrelated processes. After assessing the situation and identifying a specific threat, there are often two alternative decision making outcomes. In the active approach, the forecaster may determine that the threat is significant, and as such, they may decide to issue a product, such as a watch or a warning for a particular threat type. In the second type of approach, the forecaster may determine that the threat is not significant at that point in time, and the action would be to wait for more information or to turn his attention elsewhere.

Focus group participants discussed the importance of threshold-based assessment, or as those in the human factors profession might refer to it, *recognition-primed decision making*. Recognition-primed decisions are those decisions made in response to uncertain and often short time-frame situations, and decision makers select the first functional decision, even if it is not the optimal solution (Klein, 1989). Hoffman and Coffey (2004) had found that pattern recognition and mental modeling were deeply ingrained into the weather forecast process. Likewise, focus group participants frequently mentioned pattern recognition and threshold detection during situation assessment. Thresholds were often discussed in terms of time (e.g. deciding whether a flash flood would occur within six hours of the present time or not) or in terms of modeled parameters (e.g. detecting if the flash flood guidance exceeded the amount needed for flash flooding in a specific region). However, some participants recognized problems with basing forecasts solely on threshold detection. While basing decisions on recommended thresholds led to a sense of certainty, such decisions could lead to tradeoffs with respect to forecast verification scores, which may affect mental models used in future forecasts.

Uncertainty management also influenced the action selection stage. Here, themes related to *background experience and training* and *risk tolerance*. *Risk tolerance* refers to the degree of risk that a decision maker is willing to accept, and has been discussed as a major factor in decision making in the literature. Participants shared examples of times when their or a colleague's forecasting practices changed as a result of a previous negative outcome. False alarms (issuing a forecast for a weather event that fails to materialize) and misses (failing to issue a forecast for a weather event that does materialize) were described as influential events that sometimes led to readjustment of risk tolerance. One participant discussed the case of a coworker who had been "burned" by a missed event, and afterwards issued warnings more liberally in order to minimize the odds of missing another event. Situations like this appeared to develop forecasters' mental models, cultivating the information source held within *background experience and training*.

When speaking about making judgments about the need for a warning, focus group participants discussed the connection between their background knowledge and their ability to recognize patterns and detect environmental anomalies. Experience was discussed in terms of experiential learning as well as professional training. Comments related to training focused on formal education, often aimed at developing forecasting skill through practice and putting institutional policy into practice. Comments related to background experience were similar, but instead referred to knowledge developed at an individual level; an example of this was local knowledge built up over a period of time at a specific forecast office. Underlying the experience discussion was the concept of technology transfer—the handoff of technology from a developer to an end user. Participants revealed that their acceptance of new information sources (e.g. new decision support tools or models) not only influenced how they arrived at a decision, but their action choice, as well. Discussions suggested that tool acceptance in HWT-Hydro, specifically, may have been a function of product skill, user calibration, and availability of instruction.

Sources of Uncertainty and Challenges in the Testbed. Several sources of uncertainty posed challenges to testbed participants. The analysis revealed four primary challenges affecting decision making in the testbed: differences in participant background, a lack of information, geographic scale complications, and workstation setup issues. Much discussion centered on differences between forecasting policies that differ between offices around the country. In one situation during HWT-Hydro, a testbed participant issued a flash flood warning for a particular region, but the local WFO did not issue the same type of warning. The following day, reports of water over roads and other flooding-related outcomes were received. The participant felt justified in her original forecast and attributed the difference to variations in office policies: "I have seen where it comes back to the definition of a flash flood… at my office, we have our set definitions… we try to quantify it in terms of depth of water, moving water… In my mind, flash flood is different to what [the local office was] thinking a flash flood is."

Forecasters also identified uncertainty associated with information sources, reflecting issues with technology comprehension. Specifically, forecasters were challenged on three fronts: they lacked several traditional decision support tools, they lacked relevant local knowledge for much of the United States, and despite training, the participants were largely novice users of the experimental guidance products. Due to technical limitations imposed by the required systems for displaying the experimental guidance products, testbed workstations were unable to provide several types of commonly used information sources. Participants were able to access some of these sources through an internet browser, but this limited direct comparison of data types. In addition to this, participants did not have pre-existing mental models related to the experimental guidance tools, so it took time each week for many of the participants to become accustomed to using the model outputs as decision support in real-time forecasts.

Lack of local knowledge was a particularly difficult challenge for many participants, and was perhaps the leading factor associated with increased uncertainty. In the words of one participant, this situation "reinforced to me what local knowledge of your forecaster does for you… [we] bounced around different parts of the country, and the way things respond, changed quite a bit. So, the local knowledge is key." Participants attempted to improve local knowledge when possible, often by reviewing maps and by searching for images of the local terrain on the internet. However, this was only nominally able to resolve uncertainty. A similar challenge related to the expansive geographic scales of forecasts and resulted from the experimental design. Not only were participants asked to forecast over unfamiliar regions, but also the size of the geographic domain was much larger than the typical areal extent that a WFO forecaster would have responsibility over. This was a challenge both from a workload perspective and from a situation assessment perspective.

The final type of challenge related to workstation customization. Although participants were able to set up their workstations and displays according to their preferences, one participant stated, "I felt like I was borrowing someone else's tools," a sentiment that was echoed by others. While this may have not been a direct source of uncertainty, rapidly adjusting to a new display set up likely did not help to facilitate the uncertainty management process.

Coping Tactics. Several of the focus group questions probed participants for individual and group experiences related to coping with uncertainty. Confirming Lipshitz and Strauss's (1997) R.A.W.F.S. heuristic, focus group members largely agreed with regard to management tactics. The R.A.W.F.S. heuristic presumes that situation assessment is an adaptive, iterative process involving recognition-priming, assumption-based reasoning, and action choice evaluation. During the focus groups, conversations revealed that when faced with uncertainty, forecasters did attempt to use reduction methods to diminish its effects. Reduction techniques were the most frequently cited coping tactic, as shown in Figure 19. Participants also identified techniques that acknowledged uncertainty when further reduction was not possible; these techniques were often policy-oriented from an office- or NWS-wide context.

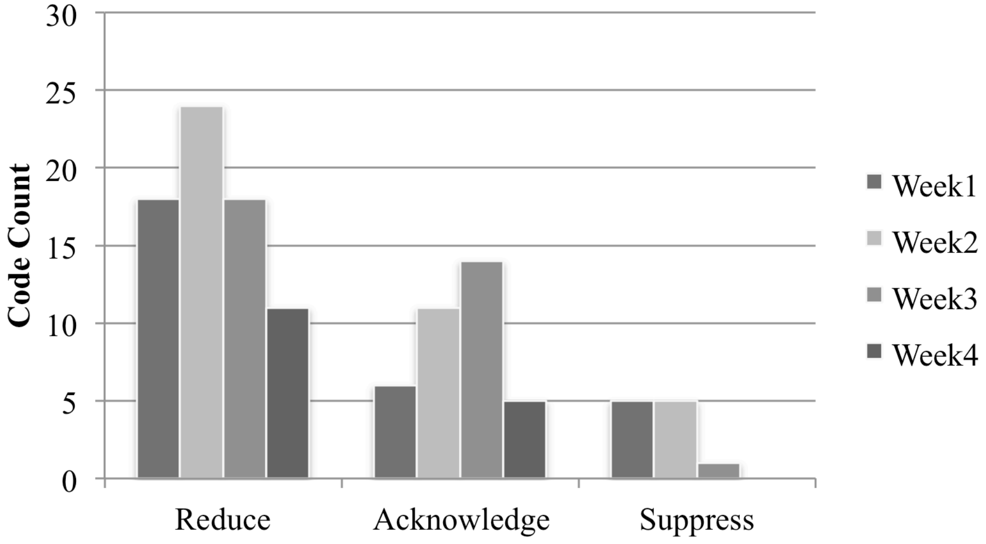

Figure 19. Count per code per week demonstrating tactics for coping with uncertainty during HWT-Hydro, shown in the R.A.W.F.S. framework proposed by Lipshitz and Strauss (1997)

Finally, suppression was discussed as a technique that was typically undesirable. Many of the identified tactics aligned with those discussed by Lipshitz and Strauss (1997) but with several differences that reveal management practices specific to the flash flood forecasting domain.

Reducing Uncertainty. In a forecasting session, uncertainty can arise from insufficient information. In each of the four focus groups, participants discussed multiple reduction tactics in the situation assessment process (60.2% of coded uncertainty management comments). Forecasters primarily cited reducing uncertainty through information-seeking activities (35.2% of all reduction-oriented comments). In flash flood watch and warning decisions, forecasters stated that they sought two types of data in particular: information that reduced uncertainty about the environmental state and information that reduced uncertainty about interpreting guidance tools. Lack of information, one of the greatest sources of uncertainty, was often exacerbated by lack of local knowledge. As one participant in the first focus group explained:

> "That was what I struggled most with… especially out west, just my unfamiliarity, as in, where's this water going to go? I looked at the, called [the town] up on Google Earth, zoomed in… this kind of looks like it could be, oh, you know, affected by this flooding up on the hilltop there, and most of the time it turned out we were wrong."

As evidenced by this forecaster's experience, reducing uncertainty by seeking additional information did not always result in a successful forecast. However, focus group participants cited a number of information sources that helped to reduce uncertainty by filling in pieces of the puzzle. Apart from geographic and topographic data, participants actively sought information about model bias adjustments (mathematical corrections to align simulated predictions with real-world observations), environmental observations (e.g. rain gauge measures and warm cloud depths), and temporal measures (e.g. mean storm motion), among others. In addition, many of the experimental tools introduced uncertainty into the decision process, and during the testbed, forecasters found themselves seeking information about the new decision aids. Furthermore, information was acquired through communication with other forecasters in the testbed as well as in briefings given by participants in a separate testbed, a behavior also discussed by Morss and Ralph (2007). Each day, HWT-Hydro participants would participate in a conference call with the other testbed participants, and the discussions would help to identify areas of concern for flash flooding across the country.

While actively searching for new information to construct and update mental models during situation assessment, forecasters also attempted to use their mental models to simulate possible outcomes, referred to as assumption-based reasoning by Lipshitz and Strauss (1997). One forecaster illustrated this tactic in a story from a testbed forecasting session, in which he issued an experimental warning for flash floods in his hometown which failed to verify. He had a mental model in place that included detailed local knowledge. He had been exceptionally confident while creating the forecast. When asked about his high level of confidence, he replied:

> "Well, the weather didn't do what I thought it was going to! No, I was looking at a downstream precipitation with the return periods on in an area and it was going to reach a flash flood guidance, which in my mind was lower than what the [River Forecast Center] had, and, uh, apparently, well, it caught the south side of town, but it didn't cause a problem. But, if it would have continued along the path it had been before it died out… it hit the county border and diminished as it got into the town… or otherwise I think it would have worked out fine."

Despite knowledge of the local region and mental models refined through years of experience, unverified forecasts do occur. Although this introduces questions related to forecast "goodness," it also exemplifies the assumption-based reasoning tactics for uncertainty reduction.

Lastly, focus groups touched on two of Lipshitz and Strauss's (1997) remaining reduction tactics: waiting (9.86% of reduction comments) and following norms of practice (11.3% of reduction comments). Although not mentioned as frequently as practices related to information acquisition or assumption-based reasoning, it appeared that several participants relied on operational standards of practice to guide decisions. This is not at all surprising; in operational settings, forecasters operate under strict directives on how and why weather products may be issued. Participants referenced organizational policies issued by the NWS, and they also mentioned office-to-office policies that affect their decision processes. In one instance, a participant stated that if they ever saw a flash flood guidance ratio reach 150%, they would immediately put out a flash flood warning, even if other information sources disagreed. In the case where the other reduction tactics were not sufficient to improve an individual's confidence past the threshold for action, several participants stated that they would wait for the situation to unfold further.

Acknowledging Uncertainty. In the taxonomy presented by Lipshitz and Strauss (1997), decision makers use acknowledgement tactics to cope with uncertainty when reduction is not possible. Out of all focus group comments coded into uncertainty management categories, 30.5% captured behaviors or beliefs related to acknowledgement tactics. Acknowledgement codes captured tactics that were typically organizational-level and policy-oriented, occurring outside the immediate forecasting timeframe. These tactics often complemented reduction tactics that were often used by individuals in the timeframe immediately surrounding the weather event in question. During the thematic analysis, coded comments aligned with tactics observed by Lipshitz and Strauss (1997): preemption (36.1% of acknowledgement-oriented comments), improving readiness (47.2% of all acknowledgement-oriented comments), preparing contingencies (8.33% of all acknowledgement-oriented comments), and consideration of action pros and cons (8.33% of all acknowledgement-oriented comments).

In the weather enterprise, preemptive action and improving readiness to negative outcomes initially appear similar, but the focus group discussions revealed several distinctions. Preempting uncertainty was defined in terms of preparing responses to anticipated events, whereas improving readiness was inherently associated with unanticipated events. Improving readiness, which Lipshitz and Strauss (1997) define as developing "a general capability to respond to unanticipated negative developments," was interpreted to refer to development of organizational policies to support uncertainty management and minimize negative outcomes. Elements of readiness included regular forecaster training, allowing individuals to customize workstations, development of new decision support tools to overcome regional uncertainties, and setting policies in place to minimize risk.

Participants discussed several degrees of preemption, ranging from testbed-specific behaviors and operational practices at the individual and organizational levels. Only able to partially reduce uncertainty associated with unfamiliar guidance product interpretation, several forecasters recognized that they adjusted their warning thresholds to reduce missing flash flood events. In the words of one participant, "I lowered my threshold. So, I was issuing more products than I normally would back home." When asked about uncertain situations in an operational setting, discussions revealed the influence of socio-geographic factors on forecasting thresholds. Despite the overarching mission to forecast weather regardless of location or anticipated impact, infrastructure and sociological factors appeared to affect some decisions. Summarized well by one participant:

> "As that level of severity increases, especially over an area where you know is, is a wilderness area, you kind of hit that threshold and say, "boom, I'm [going to] issue at this point." Whereas, like, that threshold is [going to] be a lot lower over a metropolitan area. You're [going to] be jumping on it right away. There are a lot of other factors that are going into effect."

In this example of preemption, the forecaster adjusted her threshold to cope with the situational uncertainty. While anticipating some type of negative outcome, the forecaster still recognized uncertainty surrounding the level of environmental response. This type of response was similar to policy-centered discussions; one participant discussed a forecasting policy unique to her home office that differed philosophically to other forecast offices. According to the forecaster, her home office was not willing to accept uncertainties associated with local infrastructure, such as clogged drainage systems causing localized flooding. Her office had developed a policy to issue a specialized statement to advise residents to expect heavy rainfall and localized ground-based effects, but avoided issuing location-specific warnings about flooding. This policy, leading to forecasts based heavily on rainfall observations, removed some situational uncertainty while providing the public with actionable information.

Suppressing Uncertainty. Although mentioned infrequently during focus group discussions, suppression did surface as a management tactic (9.32% of all uncertainty management comments). Suppression tactics are characterized as activities that involve denial or unfounded rationalization in order to overcome stalled decision making

(Lipshitz & Strauss, 1997). It is not uncommon to hear forecasters discussing decisions based upon “intuition.” While challenging, distinguishing unfounded “intuition” from assumption-based mental simulations has implications on understanding uncertainty management; the distinction may lie in whether or not a forecaster possesses an adequate mental model for the situation at hand. Three participants relayed stories about intuition-based forecasts made during the testbed. In each case, it was understood that the uncertainty associated with the experimental guidance products and unfamiliar geographic domains overwhelmed the forecasters, leading them to base decisions on an insufficient level of situation awareness.

Likewise, some participants acknowledged ignoring situational uncertainty on occasion. Ignoring uncertainty was discussed particularly in the context of testbed forecasting activities. When uncertainty was high, especially when it arose from lack of local knowledge, some participants built confidence from insufficient situational assessments. While these decisions were partially informed, such comments revealed the occasional instance of acting with certainty without seeking additional data. In one such case, one forecaster stated that they weren’t familiar with weather patterns in the eastern United States, so when they were asked to forecast there during HWT-Hydro, she used QPE guidance “as sort of gospel truth.” Similarly, several participants acknowledged that gambles have a role in uncertainty management. A different forecaster, citing geographic unfamiliarity, stated that when decision aids presented guidance values near a threshold, he “tended to side with the lower [values]. Just ‘cause.” Without being able to find information to reduce uncertainty and support a deeper analysis of guidance products, forecasters risk increasing negative outcomes, such as misses and false alarm forecasts.

Impact Characterizations in Forecasting Communications. During the testbed, participants were instructed to include a new type of uncertainty estimate in their experimental forecasts. With each experimental watch and warning, participants assigned a probability of a particular magnitude to their forecasts. One aim of these impact characterizations was to communicate uncertainty to forecast end users. In addition to exploring tactics for managing uncertainty, several focus group questions probed participants for feedback on positive and negative aspects of shifting towards such a paradigm.

Findings on Magnitude Attributes. Several themes emerged from the focus group responses regarding the inclusion of a magnitude estimate. Overall, the experimental requirement to include a magnitude estimate was seen as a positive addition to forecast products. Participants generally expressed a desire to have the ability to issue products with standardized text reflecting threat level in their operational office settings.

Some forecasters discussed their wishes to be able to communicate their mental model to forecast end users. Including an impact-based uncertainty statement was viewed as a means to such an end. In regard to including the magnitude and uncertainty attributes in the experimental products, one forecaster stated:

> "We kind of do that in our head. I think that's very valuable information for the public, and having this nuisance or major, we're in effect giving them that information that they would have never gotten before."

Participants also identified the need for data-driven decision support systems in operational environments. Including a magnitude estimate in a watch or warning was seen as a value-adding attribute that would help to provide actionable information that would help consumers like emergency managers to make informed decisions. In the words of one participant:

> "It gives you the ability to quantify the anecdotal information. If we're doing a decision support service brief to emergency managers, you know, on that phone call, we'll say… 'this will be a widespread, minor flood event, or… it's not going to happen everywhere, but if it does, it's going to be really bad.'"

While participants generally adopted a positive affect towards the magnitude uncertainty attributes, they did have some concerns about their design. Themes related to professional interpretation challenges, concern for members of the general public, and training issues emerged from the discussions. Some participants expressed concern that members of the public would have trouble interpreting both the probabilistic and magnitude components of the threat attributes. Furthermore, participants repeatedly commented that they would expect to see disagreement at a professional level regarding interpretation of a nuisance versus a major flood. The categorization was seen as subjective. A commonly heard comment was that what may seem like a nuisance flood from a forecasting perspective may feel like a major impact to an individual affected by it. As put by one forecaster,

> "If I get a foot of water in my basement and I'm the only one in 500 miles that did… that's a nuisance, but to me that's major."

Additional concerns tended to revolve around the lack of background experience in issuing magnitude uncertainty attributes. Although some participants stated that they regularly considered threat levels and uncertainty when issuing forecasts, comments from other participants revealed that issuing the experimental attributes created a substantial challenge for some. This may be due to a lack of probabilistic flash flood forecasting in operations and only a short training session on issuing products with the experimental attributes prior to the testbed.

Findings on Probabilistic Information. When asked specifically about the role of probabilities and factors that influence them in flash flood forecasting, positive-affect themes included mental model building, decision support services, and improved forecaster behavior. Almost as a whole, participants commented that they often considered probabilistic information during operational forecasting. While flash flood forecasting is not currently issued probabilistically, some participants suggested that they regularly consider the probability of a threat when before deciding to issue a watch or warning. This is in line with the National Weather Service's Directive 10-922, which creates thresholds for uncertainty that a forecaster must reach before issuing a watch or warning (National Weather Service, 2011). The Directive, which requires that there must be a 50-80% chance of flash flooding before issuing a flash flood watch, among other requirements before issuing a flash flood warning, may have led to some bias in the experimental watch and warning products. When asked to give an example of how a forecaster considered probabilistic information in forecasting, one participant responded:

> “In issuing a product, [I] will always consider probabilities, because innately in the directive… you must have an eighty percent confidence for something in a warning, or a fifty percent confidence in it happening for a watch. So that’s something you’re always considering.”

Another recurring theme focused on how the experimental threat attributes assisted the participants in making fewer hedged forecasts. Hedging, defined by Murphy (1978) as a forecast in which there is a “difference between a forecaster’s judgment and his forecast.” Some HWT-Hydro participants felt that by being forced to consider the uncertainty and assign a magnitude uncertainty attribute to each watch and warning, their ability to hedge was reduced; generally, this was a desirable outcome.

Study II Discussion and Recommendations

The testbed study was the first to incorporate uncertainty attributes into the forecast decision making process. While requiring further research to determine appropriately designed experimental threat attributes, they show promise in their ability to communicate forecaster SA to end users. When asked whether or not the magnitude and probabilistic categories were appropriate, participants felt that the probabilistic levels were fine for their current forecasting skill level when using the experimental FLASH tools, but it could be useful to have a scale with smaller intervals for operational forecasting. To address this concern, future iterations of the hydrology testbed experiment will allow forecasters to select probabilities at thresholds spaced one percentage point apart.

From an evaluation standpoint, it was very difficult to separate probability from magnitude in the discussion. Both were so closely linked that it was difficult to get a clear picture of how probability and magnitude were chosen separately. In addition, probability thresholds for major and nuisance flooding changed based on environment and socio-geographic constructs. Participants discussed differences in probabilistic thresholds that they needed to reach in order to issue warnings over rural and urban areas.

Based on responses from the focus groups, three recommendations were developed for the future of flash flood forecasting and decision making research. With regard to the development of impact- and uncertainty-based forecast products, participants expressed the need for consistent, actionable terminology, and a standardized scale for flood threat level. Participants pointed out that terminology often varies when forecasting for river floods, areal floods, and flash floods. Although the HWT-Hydro focused entirely on flash floods, the participants generally worked in professional roles that required them to issue warnings for other types of flood threats as well. A unified flood forecasting system requires consistent terminology to facilitate communication between actors in the weather response system.

Testbed participants also indicated that the term “nuisance flooding” was difficult to define from a scientific and a social perspective. There is a great need for future research to address best practices with regard to what type and quantity of information should be shared with different types of forecast consumers. For example, an emergency manager may be able to make a more informed decision after receiving a magnitude uncertainty attribute issued alongside a warning polygon, but an individual in a different role may interpret this type of information differently.

Although some forecasters stated that they do discuss potential impacts with forecast end users, there is currently no standardized method of communicating such risks to forecast end users. Initiatives such as Impact-Based Warnings (IBW) have experimented with the design of text-based forecast products that contain information related to potential impacts. An evaluation of IBWs for tornado threats revealed that up to a certain threshold, including possible impacts in the text product increased the likelihood that an individual would take protective action (Ripberger, Silva, Jenkins-Smith, & James, 2014). Furthermore, following a severe thunderstorm in Abilene, Texas in which an IBW was issued operationally, Guerrero, Myers, Lyons, Dunn, and Johnson (2015) found that the additional impacts-oriented text gave members of the public actionable information that lessened confusion and clarified the level of risk.

Lastly, future work is needed to develop a scale for flash flood forecasting impacts. Unlike the Enhanced Fujita Scale for tornado threats, there is no scale available for use by National Weather Service forecasters for communicating flash flood threat level. The nuisance and major flood categorizations used in the magnitude attributes in HWT-Hydro attempted to provide a basic structure for flood threat. However, additional research into scientifically and socially appropriate threat levels would be of great benefit to the forecasting community and society at large.

General Discussion

The purpose of this study was twofold. First, we aimed to identify SA requirements for flash flood forecasting and their evolution over temporal and environmental activity scales. Second, we used the mixed methods approach to explore situation assessment during flash flood forecasting. In the first study, we used a quantitative approach to assess SA requirements for flash flood warning decisions. In the second study, we employed a qualitative, focus group approach in order to categorize tactics regularly used for managing uncertainty. Originally, we hypothesized that SA requirements would differ between the watch and warning issuance timeframes, as measured by the time and frequency of guidance usage. We also expected to observe differences in SA requirements between forecast periods with varying levels of environmental activity. In relation to understanding uncertainty management in situation assessment, we hypothesized that tactics used by forecasters would align with the R.A.W.F.S. heuristic described by Lipshitz and Strauss (1997).

The mixed methods approach may have been a novel approach to these research questions, but comparing between the two datasets allowed us to make new inferences about the relationship among uncertainty, situation assessment, and decision making in weather forecasting. Alone, the focus groups and thematic analysis add to existing knowledge about decision making, situation assessment, and uncertainty management. In combination with the quantitative SA requirements analysis, though, the qualitative data is enhanced with empirical evidence. Viewed together, we are able to draw conclusions about the role of technology in uncertainty management and situation assessment.

Situation Assessment in Weather Forecasting

Situation assessment in weather forecasting is a dynamic process that is influenced by individual, organizational, and technological factors. The thematic analysis grounded the SA requirements analysis by contextualizing situation assessment in the broader scope of the entire forecast decision making process. Focus group participants described weather prediction as a process in which forecasters attempt to understand the environmental situation by assessing information sources, initially in order to understand the broad context and then through focused attention on at-risk geographic areas. These activities can be viewed as part of the situation assessment process that precedes the action selection and implementation process. Action selection and situation assessment shared several aspects in common; both processes involved recognition-primed decision making and were influenced by individual factors such as background experience and risk tolerance.

This description of the forecast decision making process aligned with accounts found elsewhere in the literature. Although not framed in terms of situation awareness, Morss and Ralph (2007) presented a procedural model of forecaster decision making and suggested that forecasters assimilate information gained from individual knowledge, model guidance, observational data, and interpersonal communication as inputs into the forecast decision. Similarly, Doswell (2004) framed forecasting as a cycle of diagnosis and prognosis. In relation to Endsley's (1995c) model of situation awareness, diagnosis may be equivalent to Level 1 (perception) and Level 2 (comprehension) SA, while prognosis may be similar to Level 3 SA (projection). From a macrocognitive perspective, Trafton and Hoffman (2007) suggested that forecasting begins with an action queue, using iterative situation assessment and recognition-primed decision making to build mental models and situation awareness, culminating in action selection.

The thematic analysis findings not only corresponded with these representations, but they also provided evidence that bridged procedural and macrocognitive models of weather forecasting. In addition to gathering qualitative descriptions of the forecasting process, one of the study's main contributions was the identification of situation awareness requirements for dynamic comprehension and projection of flash flooding situations. During the testbed, we observed that forecasters relied on different guidance products between watch and warning issuance stages, which confirmed the timeframe hypothesis. More surprisingly, the observations confirmed differences in SA requirements as environmental activity level increased. This deferral of situation assessment from high-activity days to low-activity days suggests that SA requirements may be satisfied over a long-term forecasting period, such as several days or even weeks.

Testbed participants did not possess local knowledge for all the geographic locations they issued forecasts over during the study, and the time study results indicated that they consulted more products related to understanding geography and initial conditions on low-activity days. It is possible that SA requirements, such as information on local geography, can be deferred to low-activity days. Going forward, this may inform the design of additional forecast decision making studies that are involved in studying guidance usage.

Uncertainty Management Techniques

Despite being a useful construct in many domains, some accounts of situation awareness and situation assessment are limited when they are applied to complex decision making domains that involve uncertainty (Minotra & Burns, 2015). In the weather forecasting domain, understanding forecaster techniques for coping with uncertainty may have theoretical implications for understanding SA, and practical implications in terms of decision support development. In the present study, we sought to identify how forecasters cope with uncertainty in situation assessment, and in turn, how such uncertainty affected the whole decision making process.

Doswell (2004) suggests that forecasters incorporate uncertainty into decision making by combining progressive, intentional, and logical analysis with the reverse: intuition. While this framework captures the broader essence of forecast decision making, the mixed methods study provided insight into the factors and processes at work within Doswell's (2004) two modes. In the thematic analysis, we examined the focus group discussions through the theoretical lens provided by the R.A.W.F.S. heuristic (Lipshitz & Strauss, 1997). The thematic analysis revealed that forecast decision making to be a function of several factors, and that forecasters manage situational uncertainty through a number of individual and organizational management tactics. Forecasters discussed employing reduction tactics on an individual and group basis, as well as suppression methods, though to a lesser degree. Organizational policy and best practices provided context for individual-level forecasting decisions, and focus group participants often framed these policies in a way that aligned with Lipshitz and Strauss's (1997) definition of uncertainty acknowledgement.

Although Lipshitz and Strauss (1997) originally discussed the R.A.W.F.S. heuristic as it related to militaristic decision making, we suggest that the heuristic may be generalizable to the complex domain of weather forecasting. Here, we found that forecasters regularly discussed using reduction tactics, including goal-directed information collection, decisions based on organizational norms, soliciting guidance from colleagues and technology, and forestalling when necessary. Interestingly, several reduction-oriented tactics align with behaviors observed in naturalistic decision making studies of weather forecasters (Kirschenbaum, 2004; Trafton, 2004; Trafton & Hoffman, 2007). In Lipshitz and Strauss's (1997) tactical framework, assumption-based reasoning refers to uncertainty reduction via use of a mental model based on constrained beliefs and evidence related to the situation. Following the testbed, focus group participants were aware of practicing such behavior while forecasting.

Several mechanisms and factors may relate to forecaster cognition and uncertainty management. Assumption-based reasoning is linked to the creation and implementation of a mental model. Trafton (2004) suggested that weather forecasters develop qualitative mental models that permit the forecaster to draw inferences dynamically about the environment. In the present study, the quantitative results reflected the differences in information sources needed to build SA and a mental model of the situation. Trafton and Hoffman (2007) found that forecasters develop their mental models by using spatial transformations to synthesize spatial-temporal information into a refined understanding of the situation; working with forecasters, they identified that the most frequent type of spatial transformation was comparison between information sources. One of the emergent themes from the thematic analysis focused on the centrality of comparison in understanding the broader situation and specific threats.

It is also been suggested that expertise plays a large role in uncertainty management. In a study of military tactical commanders, St John, Callan, Proctor, and Holste (2000) varied situational uncertainty and found that inexperienced participants employed a "wait-and-see" tactic more often than experienced participants. While focus group discussions did reflect that forestalling a decision was an accepted tactic for forecasters, it was less frequently mentioned that other tactics. Participants in the present study were balanced in terms of expertise, but naturalistic decision making accounts may provide insight into situation assessment through explanations of recognition-primed decision making. In its basic form, pattern recognition and recognition-primed decisions are closely aligned, but as situational uncertainty increases, decision makers must rely upon mental models and mental simulations. Expertise governs a decision maker's ability to perform successfully in these activities (Lipshitz et al., 2001). Participants in the present study possessed relatively equal levels of forecasting experience and exposure to the experimental decision aids, which may explain infrequent references to forestalling tactics.

Theoretical Contributions to Understanding Uncertainty and SA

The mixed methods study produced several findings that extend Endsley's (1995c) Model of SA to decision making under uncertainty. Some findings directly align with several components of the model, whereas other findings provide insight into less-explained aspects of situation assessment.

The focus group discussions reflected the influences of background experience, system design, and risk tolerance on forecasting. Endsley (1995c) proposed that SA consists of three levels (perception, comprehension, and projection) and that SA is influenced by task/system factors (e.g. interface design, stress, workload, and automation) and individual factors (e.g. goals and preconceptions, expertise, and long term memory). Forecasters described their assessment and prediction process in alignment with the three levels of SA. Perception occurred as forecasters sought information and consulted guidance products, while comparing between and within guidance products developed comprehension. Confidence in projections increased as situational uncertainty decreased.

Endsley (1995c) framed SA as an in-the-head model of the current situation, which when compared to a global mental model, can facilitate recognition-primed decision making. Endsley (1995c) also suggested that, over time, operators build and refine new mental models as SA is developed in new contexts, and that decision makers actively partake in goal-directed information assessment. During the experimental watch and warning issuance activity, forecaster behavior not only reflected SA requirements, but also provided additional evidence to support the role of goals in SA. Operator goals, such as "determine if risk is high enough for a warning," are part of top-down processing, in which goals and preconceptions direct the forecaster's attention when searching for information to reduce uncertainty and build SA. Bottom-up processing was also discussed in the focus groups; as forecasters detected anomalies in the environmental activity, such observations would in turn guide information seeking. The quantitative results also support this; guidance products were often used in comparison activities, one of the most common spatial transformations (Kirschenbaum, 2004; Trafton & Hoffman, 2007). For example, a forecaster may have observed that the QPE-to-FFG ratio levels exceeded 150% in a certain region, which then prompted them to assess other guidance products over that same region.

While many of the present study's findings concurred with existing theory, they also extend current explanations of SA and the forecasting process under uncertainty. In the original model, Endsley (1995c) conceptually acknowledged that uncertainty affected decision maker confidence, which in turn could affect decision outcomes. Based on the mixed methods findings, we suggest that forecasters cope with uncertainty through reduction, acknowledgement, and suppression techniques, as framed in the R.A.W.F.S. heuristic by Lipshitz and Strauss (1997). When time permitted, forecasters actively sought additional information to reduce their uncertainty. Furthermore, discussions revealed that organizational policies were often in place within operational offices to reduce the effects of potentially negative outcomes related to decisions made under uncertainty. However, when uncertainty existed even after reduction and acknowledgement, discussions revealed that suppression did occur. Such tactics allowed forecasters to build a dynamic situational model that accounted for potential alternative scenarios as well as the most likely outcome. Indeed, forecasters appeared to be most concerned with the effects of uncertainty on their projections (Level 3 SA) and their ability to comprehend the environmental and atmospheric situation (Level 2 SA).

Limitations

This work resulted in several insights into SA and decision making under uncertainty in flash flood forecasting. Nevertheless, a number of limitations exist that must be considered when drawing conclusions from the data. The focus group sample size was smaller than has been recommended in the literature (Caplan, 1990), with only four participants at most per group. One of the main concerns with a small sample is group proclivity towards a single, dominant opinion, leading to difficulties in stimulating new discussion. As it were, conversational themes identified in the thematic analysis aligned with theoretical accounts of uncertainty management and forecast decision making, so we believe that the sample size was sufficient.

In a departure from other studies related to SA and decision making, we did not look at each participant's level of SA, but instead at their information requirements on the assumption that they were building SA. While we did not assess decision making using traditional Naturalistic Decision Making methods, we were concerned with observing forecasters "in situ" and understanding forecaster behavior in their own words. In Study I, we assumed that the presence of a guidance product on the computer screen equated to it being used by the forecaster. This assumption meant that guidance products were recorded even if the forecaster did not consciously extract information from them. However, we hypothesized that a tool's presence in the periphery may have subtly affected judgment. While the method did not produce the same degree of accuracy as a method like eye tracking would have, this was tempered by measuring the time each participant spent viewing each of the products.

More critically, the Study I analysis was limited in that the several of the sampled videos had visual quality issues. In the majority of the screen recordings taken by participants working at the dual-monitor workstations, the software only produced interpretable recordings of one of the two monitors. As a result, although participants viewed guidance products on both monitors, we intentionally sampled recordings in which information was not only legible, but in which most of the interaction occurred on the visible portion of the screen. A similar issue that we were not able to work around was that participants regularly brought in tablet computers and personal laptops, which they used to consult unofficial forecasting guidance products over the internet.

Summary

The primary goal of this study was to explore the relationship among situation assessment, uncertainty management, and decision support tool usage in weather forecasting. Through the mixed methods analysis, we were able to provide examples of certain links between theoretical accounts of SA and of uncertainty management. The quantitative results supported the hypothesis that SA requirements differ for decisions in the watch and warning timeframes as well as at increasing levels of environmental activity. Adding to previous explanations of forecast guidance usage in flash flood forecasting, the present findings indicate that hydrology-based guidance products may provide information, that when used in combination with other decision support tools, can improve forecaster SA. Focus group discussions with professional forecasters revealed that uncertainty management techniques identified by NDM studies in other domains are also practiced in weather forecasting.

Uncertainty management and risk reduction in the weather domain has previously been attributed to emergency managers' actions (Morss et al. 2015). This study demonstrates that risk reduction is part of the weather forecaster's purview as well. It is evident that uncertainty management is an integral part of situation assessment, and that comprehension of uncertainty in the forecast process can improve overall SA. In order to further assess the relationship between forecast uncertainty and SA, future work should assess SA levels of forecasters while using guidance products. Understanding the effects of uncertainty on SA and forecast decision outcomes will not only illuminate how uncertainty propagates throughout the weather enterprise, but it will also contribute increased knowledge into SA development among individuals in complex systems.

Chapter 5: Automation and Situation Awareness in Flash Flood Forecasting

Introduction

During weather prediction activities, forecasters actively seek information through top-down and bottom-up processes in order to establish situation awareness (SA) (Hoffman & Coffey, 2004; Trafton & Hoffman, 2007). As decision support tools become more complex, it will be important to consider their ability to combat information overload while improving forecast lead-time and decision making. Recent efforts have studied algorithms to automate part of the forecasting process; some researchers have proposed the development of weather forecasting recommender systems, a guidance product that would create an initial threat polygon based on predictions from a collection of weather prediction models (Karstens et al., 2015).

Until recently, recommender systems have had relatively little attention in weather prediction. However, they have had more traction in commercial domains, such as e-commerce and tourism (Braunhofer, Elahi, Ricci, & Schievenin, 2013; Burke, 2002). In these settings, recommender systems use prediction algorithms to classify items then "recommend" them to potential consumers. These algorithms can be based on attributes including user demographics, preferences, or through collaborative filtering between the system and the user (Burke, 2002). Recommender systems reduce a large amount of data based on user preferences or contextual information, which may help to improve information overload during the decision making process. In weather forecasting recommender systems, the intention is that the system would essentially automate the situation assessment process. The model-based algorithm would automate a "first pass" through situation assessment. Like recommender systems in commercial applications, forecast recommenders are intended to reduce forecaster workload while improving lead time and situation awareness (Karstens et al., 2015). Although Karstens et al. (2015) found that early designs of severe hail recommenders did not significantly reduce the amount of time it took to issue a warning, it was hypothesized that recommenders played a role in the decision making process.

Previous research has found that an appropriate level of automation for the context in question may improve operator workload, confidence, SA, and performance (Endsley & Kiris, 1995; Parasuraman & Riley, 1997; Parasuraman & Wickens, 2008; Wickens, 2008). However, some studies have shown that a high degree of automation can lead to out-of-the-loop decision making, which can reduce an operator's situation awareness (Dao et al., 2009; Endsley & Kiris, 1995). Out-of-the-loop situations occur when an operator is removed from the decision process and must rely on external actors to make decisions, which can lead to decrements in overall awareness as well as task performance (Endsley & Kiris, 1995; Kaber & Endsley, 1997). Furthermore, Parasuraman and Riley (1997) cautioned against improper use of automation, citing instances in which performance suffered from overuse, underuse, or inappropriate use of automated systems.

The previous chapters of this dissertation have established that developing situation awareness can be affected by display attributes and uncertainty ingrained in decision support systems. In the present chapter, we extend our understanding of SA in weather forecasting further by exploring the relationship between automated decision aids and SA. The recommender algorithm design for weather forecasting was outside the scope of this study; instead, we focused on understanding how their presence acted as a mechanism for directing forecaster attention during weather prediction tasks. In the following chapter, we discuss the results from an experiment that assessed the effects of recommender use on forecaster SA in a flash flood prediction task.

The Research Questions

Automation in the workplace affects task performance and decision-making within a variety of domains (Dao et al., 2009; Endsley & Kiris, 1995). Recommender systems, a newcomer in the weather forecasting domain, may have potential to reduce time-consuming situation assessment activities within the forecasting process. While some evidence exists to suggest that early versions of recommender systems may not significantly reduce forecaster workload, we hypothesized that recommenders would affect situation assessment and forecasters' levels of SA during a flash flood forecasting task. In line with this hypothesis, the primary objective of this research was to explore the relationship between recommender usage and SA (RQ3.1).

RQ3.1: How is SA influenced by recommender automation at different processing levels during a weather forecasting task?

In order to assess SA during recommender use, we employed an eye tracking system to capture data related to participants' information-seeking behaviors. To date, the literature contains only a small number of studies that intersect eye tracking, weather forecasting, and situation awareness. As such, the secondary research aim was to evaluate the relationship between eye tracking measures and SA. In what we believe is the first reported case study that employed eye tracking to assess a weather forecaster, Bowden, Heinselman, and Kang (2016) established that eye tracking provided insight into the forecast decision process. Eye tracking has only recently been identified as a feasible method for SA assessment; existing work suggests a positive relationship between eye tracking measures, SA, and decision making (Moore & Gugerty, 2010; Sturre, Chiappe, Vu, & Strybel, 2015; van de Merwe, van Dijk, & Zon, 2012; Yu, Wang, Li, & Braithwaite, 2014). In a study of air traffic controllers, Moore and Gugerty (2010) found that participants with high levels of SA fixated on relevant areas of the information display more frequently than their counterparts with lower SA. Likewise, we aim to assess the predictive power of eye tracking measures in relation to SA in flash flood forecasting (RQ3.2).

> *RQ3.2: To what degree are eye tracking measures (total fixation duration, mean fixation time percentage, time to first fixation, and mean number of fixations) able to predict situation awareness?*

Hypotheses

The present study assessed situation awareness along five metrics: response accuracy to an SA questionnaire, evaluation time (the amount of time spent reviewing the display), mean count of eye fixations, total fixation duration per area of interest (AOI) within the display, and percentage of total fixation duration per AOI. We evaluated participants' SA based on responses to probes; from these, we determined participants' SA scores at Endsley's (1995c) three theoretical levels and as a composite score based on the mean of the sublevel scores. This resulted in four scores: $SA_{Level\ 1}$, $SA_{Level\ 2}$, $SA_{Level\ 3}$, and SA_{comp}. Based on findings from Moore and Gugerty (2010), we hypothesized that decision support automation would affect the dependent variables, as listed:

H3.1: Situation Awareness Score (SA)

H_0: $SA_{available} = SA_{unavailable}$ H_1: $SA_{available} \neq SA_{unavailable}$

H3.2: Mean task duration (t)

H_0: $t_{available} = t_{unavailable}$ H_1: $t_{available} < t_{unavailable}$

H3.3: Mean number of eye fixations (n)

H_0: $n_{available} = n_{unavailable}$ H_1: $n_{available} > n_{unavailable}$

H3.4: *Total fixation duration by AOI (F_d)*

H_0: $F_{d\text{ , available}} = F_{d\text{ , unavailable}}$ H_1: $F_{d\text{ , available}} \neq F_{d\text{ , unavailable}}$

We expected to identify a difference between automation conditions in probe accuracy. In line with findings by Endsley and Kiris (1995), it is hypothesized that Level 2 SA (comprehension) will be most affected by automation level. As one premise was that recommenders would guide forecaster attention to areas of high risk, we also hypothesized that recommenders would lead to higher scores on the Level 3 SA (projection) probes in the scenarios where recommenders were available.

In terms of task duration, we hypothesized that the availability of recommenders would lead to a reduction over the condition where recommenders were not available; this was expected partially because recommenders are designed to reduce lead-time. Karstens et al. (2015) found no significant difference between warning issuance times based on recommender presence and absence; however, they based their evaluation on polygon creation time, whereas the present study evaluated situation assessment. The present study was concerned only with situation awareness, and so removed the aspect of action performance from the experimental equation. Although current evidence does not support the suggestion that recommenders may reduce issuance time, in the present study, it was expected that presence of the recommender polygons would lead to forecasters spending less time engaging in situation assessment.

In terms of eye tracking metrics, we hypothesized that automation level would affect the number of eye fixations as well as total fixation duration within the display panel containing the recommenders. While lower task durations were anticipated in the recommender-available condition, a higher number of eye fixations were expected in the recommender-available condition. This expectation was due to the recommenders providing additional visual stimuli and thus attracting participant attention. Additionally, we hypothesized that the greatest number of eye fixations and fixation durations would occur in the quadrant of the four-panel information display that contained the recommender polygons. We also assessed first fixation time within an AOI and scanning patterns across the display, but lack hypotheses due the descriptive nature of the parameters.

Method

Participants

The sample consisted of eighteen professional forecasters recruited from Weather Forecast Offices, River Forecaster Centers, and other National Weather Service Centers in the central United States. Participants had to be 18 years or older as well as currently employed by the National Weather Service (NWS). Furthermore, participants must either have held a professional forecasting role at the time of the study or prior to it. Due to the nature of the decision aids, it was preferable, though not necessary, for forecasters to work primarily in hydrologic forecasting.

Participants represented a diverse set of roles within the National Weather Service. Of the eighteen participants, eleven held roles as active forecasters (six in general forecasting and five specifically in hydrological forecasting). The remaining seven participants held current roles as forecasting and research support staff, but had held a forecasting position within the NWS prior to the experiment. Participants had a mean of 19.3 years of professional weather forecasting experience (σ = 7.95). Some participants had less experience related to hydrologic forecasting (μ = 14.6 years, σ = 7.25); however, all but one participant had responsibility for hydrologic forecasts at some point throughout their careers. Finally, participants brought a range of geographic knowledge to the experiment; Figure 20 presents the spread of forecaster experience across the United States. Several participants had previously worked in each of the

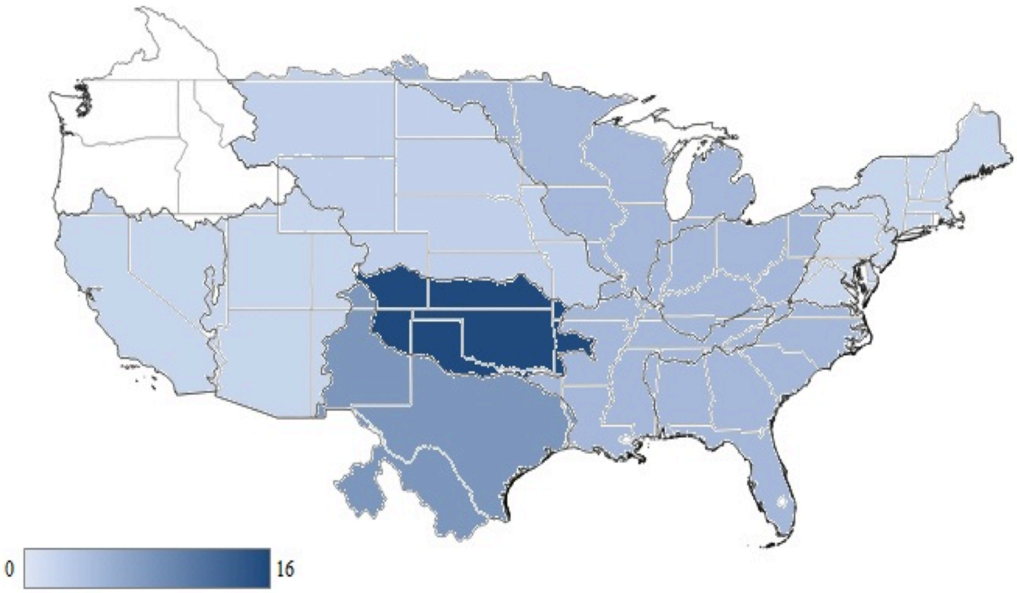

Figure 20. Map representation of the number of participants with professional forecasting experience per River Forecast Center region in the continental United States

forecast areas selected in the scenarios and so may have had a more detailed mental model of the region than other participants.

Scenario Selection

The method employed a set of scenarios displayed with and without recommenders. A set of three scenarios were selected from flash floods reported in the NOAA Storm Data publication and from cases that occurred during the 2015 Hydrometeorological Testbed Experiment. Only flash floods that occurred in May - July 2015 were selected. The following three cases were selected:

A. 31 May 2015 21:30 UTC – 1 June 2015 01:30 UTC; New Jersey
B. 12 July 2015 06:00 UTC – 10:00 UTC; Central and Southern Indiana
C. 14 July 2015 19:00 UTC – 23:00 UTC; West Virginia and Ohio Valley

Each scenario consisted of a four-hour timespan in which flooding ramped up and persisted through the end. Timeframes were chosen to coincide with the valid times of operational flash flood warnings issued by local Weather Forecast Offices. In addition to an operational warning present, historical reports from United States Geological Survey (USGS) stream gages were assessed during the scenario timeframes. Selecting scenarios that overlapped with gages that reached flood stage provided a more objective way to verify existence of a flash flood than selecting timeframes based on NWS verified storm reports alone. Furthermore, while it was required for at least one stream gage to reach flood stage in a scenario, not all gages in the region did; this allowed for an evaluation of inaccurate risk assessment.

Scenarios were divided into two-hour halves, shown in 15-minute time steps, with the exception of the radar, which updated every 2 minutes. Recommender presence was assigned randomly to each trial; half of each scenario was visualized with recommenders and the other half without. The guidance products, shown in Figure 21 and described in Table 12, were presented by running the Advanced Weather Interactive Processing System II (AWIPS-II), a computer visualization display platform, through a virtual network in order to display it on the eye tracker's monitor. Participants were not permitted to change the arrangement of the decision aids, the type of decision aids, or the color palettes of the decision aids. However, they were allowed to zoom and pan across the visualizations.

Recommender Development

Eventually, recommenders will be created through an algorithm that combines outputs of multiple tools; however, at the time of this study, such strides had not been made for flash flood recommenders. In order to test the effects of a recommender, then, a preliminary version of a recommender was created. The recommender algorithm in

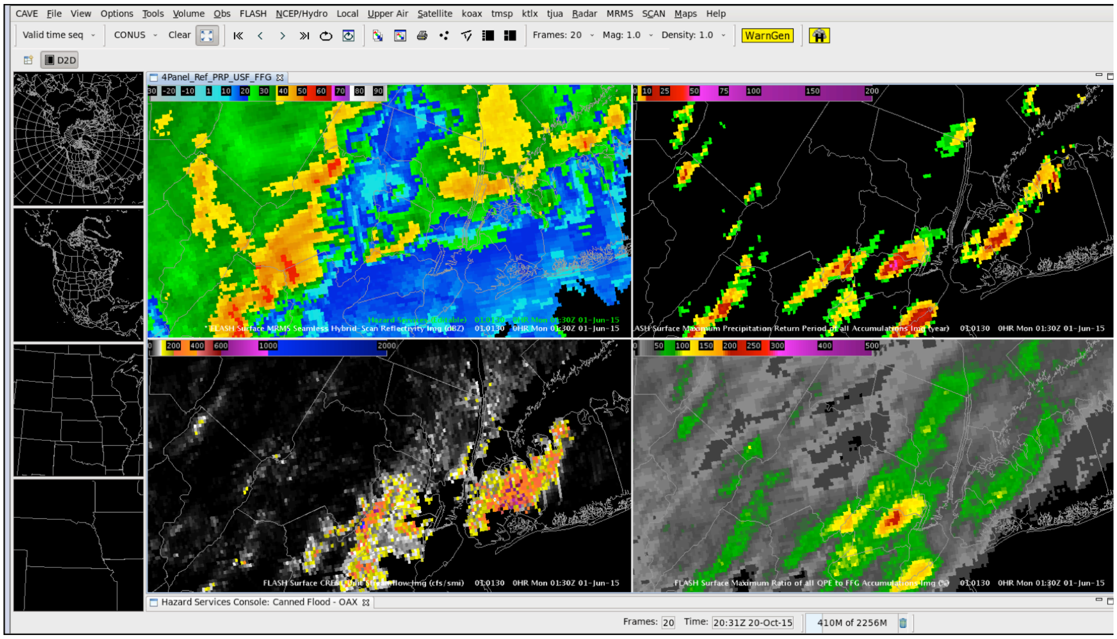
CAVE File View Options Tools Volume Obs FLASH NCEP/Hydro Local Upper Air Satellite koax tmsp ktlx tjua Radar MRMS SCAN Maps Help
Valid time seq
CONUS
Clear
Frames: 20
Mag: 1.0
Density: 1.0
WarnGen
D2D
4Panel_Ref_PRP_USF_FFG
30 -20 -10 1 10 20 30 40 50 60 70 80 90
0 10 25 50 75 100 150 200
Hazard Services (Editable) 01:0130 0HR Mon 01:30Z 01-Jun-15
FLASH Surface MRMS Seamless Hybrid-Scan Reflectivity Img (dBZ) 01:0130 0HR Mon 01:30Z 01-Jun-15
ASH Surface Maximum Precipitation Return Period of all Accumulations Img (year) 01:0130 0HR Mon 01:30Z 01-Jun-15
0 200 400 600 1000 2000
0 50 100 150 200 250 300 400 500
FLASH Surface CREST Unit Streamflow Img (cfs/smi) 01:0130 0HR Mon 01:30Z 01-Jun-15
FLASH Surface Maximum Ratio of all QPE to FFG Accumulations Img (%) 01:0130 0HR Mon 01:30Z 01-Jun-15
Hazard Services Console: Canned Flood - OAX
Frames: 20
Time: 20:31Z 20-Oct-15
410M of 2256M

Figure 21. Example of the AWIPS-II four panel interface visualizing Scenario 3

Table 12. Description of guidance products used in the present study

Decision Aid (Abbreviation)	Units	Description
CREST Unit Streamflow (USF)	$m^3s^{-1}km^{-2}$	Simulated surface water flows normalized by drainage area, selected from a span of 0.5 – 6 hours after the valid time
Precipitation Return Period (RP)	Years	Generates a return period based on precipitation rate and historical return periods. Higher return periods correspond to higher likelihood of flooding.
QPE-to-FFG Ratio (FFG)	N/A	Calculates ratio by comparing Flash Flood Guidance grid values against MRMS radar precipitation rates. Bankfull conditions may exist when the ratio exceeds 1.0.
MRMS Composite Reflectivity (MRMS)	dBZ	Mosaic of reflectivity values measured by MRMS radars across the CONUS.

the present study is based on a threshold metric: a user can select a forecasting decision aid as well as a numerical threshold based on the values the selected aid can predict. For this study, the QPE-to-FFG Ratio guidance product provided the underlying model because it was the most traditional tool for flash flood prediction used in the experiment. A threshold of 100% was selected, and the algorithm was then applied to the forecasting visualization. Represented on the map as white polygons, the recommenders were created by drawing contours around all regions contained within an area of at least 10 square kilometers of grid cells modeled at or above the threshold value.

Figure 22 shows an example of the recommenders created for the second scenario. During the experimental trials, the recommenders were always visualized in the same quadrant of the AWIPS-II display as the CREST unit streamflow

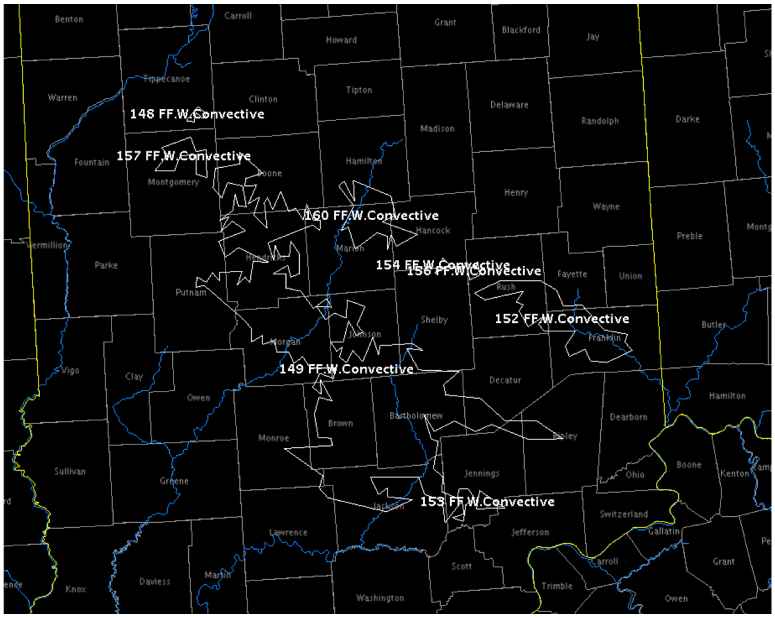

Figure 22. Recommenders in the Indiana scenario

visualization. This placement allowed for a more meaningful usage of the threshold-contouring recommender; placing the QPE-to-FFG Ratio recommender over its own base product would have merely resulted in a highlighting the regions already represented as “at-risk” by the color scale. Transposing the QPE-to-FFG recommenders into the CREST unit streamflow map theoretically would allow participants to assess the overlap in risk between the two decision aids.

Data Collection Systems

The present study used a Tobii TX300 eye tracking system to collect the physiological data. In addition to eye tracking methods, a set of probes assessed situation awareness across the three theoretical levels. A presentation technique, modeled after that used by Dao et al. (2009), was chosen for the present study. Pointing to the limitations of SAGAT in terms of working memory capacity and SPAM in terms of recall versus true SA, Dao et al. (2009) presented three probes between short simulations; the present work extended this technique to the weather forecasting domain. Probes assessed SA in alignment with the theoretical definitions proposed in the Endsley 1995 Model of SA, which frames SA in terms of perception, comprehension, and projection of environmental status into the future. In line with this framework, probes assessed participant awareness of information in the past (the information contained within the first 1.5 hours worth of data scans), the present (the final frame of data scans), and the future (expectations of flooding in the following two hours). A complete list of the probes and a copy of the scoring guide can be found in Appendix E.

Experimental Design

The study used a single-factor, within-subjects design that assessed the effects of automation use on situation awareness in flash flood forecasting. All participants received exposure to both of the treatment conditions. Scenarios were presented in a semi-random order, with probe order and automation condition also randomly presented. Automation was present in two levels (availability or unavailability of recommenders). Dependent variables were captured during the procedure: eye fixation count, total fixation duration, percentage of fixation duration within an AOI, SA score, and task duration.

Procedure

Upon arrival, participants received an explanation of the study's goals and activities. Participants read through a brief training guide on the forecasting tools, and were given an opportunity to ask questions. Following the training, participants received the prompt that they had just begun their shift and a significant rainfall event was underway. A hypothetical colleague needed them to review the prior two hours of model data and radar scans in order to identify areas of highest flash flooding risk. Participants prefaced each scenario by reading a short briefing on the status of the environment leading up to the two-hour span contained in the scenario. Briefings were selected from operationally issued heavy rainfall watches, flash flood watches/warnings, and mesoscale discussions produced by the Storm Prediction Center. An example briefing can be found in Appendix F.

Following the briefing, participants took part in the randomly presented scenarios. During each trial, participants viewed the AWIPS-II four-panel display showing each of the different weather forecasting decision aids. In the *recommenders unavailable* condition, the decision aids appeared with no alterations, but in the *recommenders available* condition, the recommenders were shown as white polygons overlaid on the upper-left quadrant of the display. Although participants could view the recommenders in the *recommenders available* condition, they were not constantly on the screen; due to technical constraints, the recommenders only appeared during the timestamp that they referenced. Thus, even in the *available* condition, participants were not always able to see the recommenders. The two experimental conditions were distinguished from each other in that in one, participants could choose to use the recommenders, whereas in the other, they were not given the option. Participants were allowed up to seven minutes to assess the state of the environment.

During each scenario, the Tobii TX300 eye tracker captured eye fixations and fixation times. In between each scenario, participants answered six probes, classified into one of the three SA processing levels (perception, comprehension, and projection).

Following the first set of probes, participants viewed the second half of the scenario with the reverse automation condition and the same instructions. At the end of the second half, participants answered another six probes. This process repeated for each of the three scenarios. Following the data collection, participants completed a background experience questionnaire and were debriefed. The experiment took 1 to 1.5 hours to complete.

Results and Analysis

At the end of each of the six scenarios, participants completed a six-item questionnaire that followed the modified SAGAT protocol; display screens were frozen and made blank while participants attempted to answer questions related to perception, comprehension, and projection based on the information they had seen. For each participant and simulation, we calculated an accuracy-based score for SA at Level 1, Level 2, Level 3, and as a mean composite of overall SA. In addition to SA scores, we also measured the task duration, or amount of time a participant spent completing each scenario. The eye tracking measures also produced a wealth of data, and provided insight into participants' behavior related to information scanning patterns.

The composite scores satisfied the assumptions of normality (W = 0.9759, p = 0.05406) and constant variance (Fligner-Killeen χ^2 = 0.2208, p = 0.6384). However, while the individual SA level scores satisfied the constant variance assumption (Fligner-Killeen $\chi^2_{\text{Level 1}} = 1.0599$, $p_{\text{Level 1}} = 0.3032$; $\chi^2_{\text{Level 2}} = 0.3781$, $p_{\text{Level 2}} = 0.5386$; $\chi^2_{\text{Level 3}} = 0.0116$, $p_{\text{Level 3}} = 0.9142$), none of the level scores satisfied normality ($W_{\text{Level 1}} = 0.8797$, $p_{\text{Level 1}} < 0.001$; $W_{\text{Level 2}} = 0.9715$, $p_{\text{Level 2}} = 0.02426$; $W_{\text{Level 3}} = 0.9449$, $p_{\text{Level 3}} < 0.001$).

In the recommenders-unavailable (control) condition, participants had a mean composite SA score of 33.59%, whereas participants had a mean composite score of 32.81% in the recommenders-available condition. Figures 23 and 24 present the distribution of SA scores (mean, Level 1, Level 2, and Level 3) between SA levels and automation condition, respectively. With a further reduced sample to ensure a balanced dataset (n = 12), a paired two-tailed t-test failed to identify a significant difference in composite score performance between the recommender-present (μ = 0.32, σ = 0.09) and recommender-absent condition (μ = 0.31, σ = 0.11), t(11) = 0.22, *p* = 0.83. While the composite (mean) SA score measure was normally distributed, the individual level scores were not. Accordingly, the Wilcoxon Rank-Sum test did not find any significant differences in performance between recommender conditions in the Level 1 score (W = 1381.5, *p* = 0.8498), Level 2 score (W = 1441.6, *p* = 0.5621), or Level 3 score (W = 1357.5, *p* = 0.9741). This suggests that recommender condition neither significantly affected a forecaster's overall nor sublevel SA.

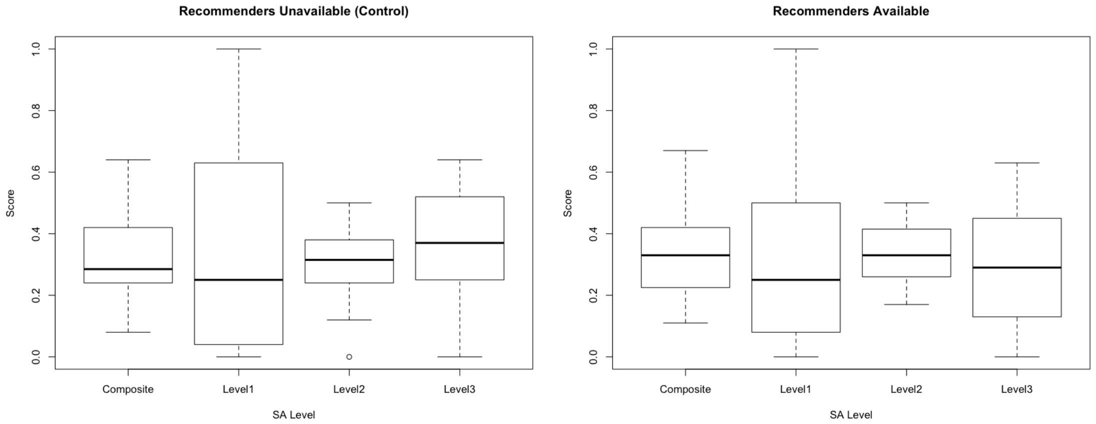

Figure 23. Boxplots showing the distribution of SA scores across performance levels and between automation levels

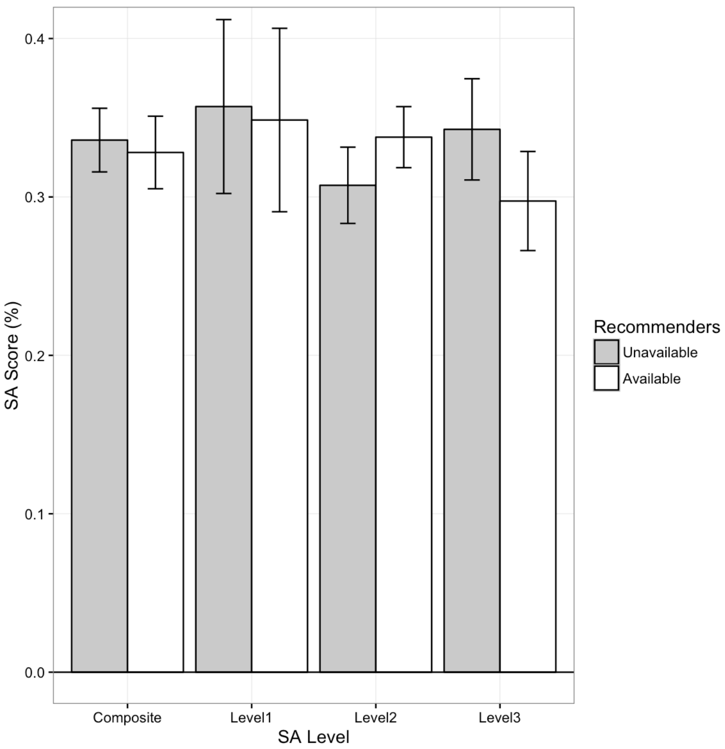

Figure 24. Mean SA score by level and condition, with standard error bars

Task Duration

To address the task duration hypothesis, which posited that task duration in the experimental condition would differ from that of the control condition, we compared the mean task durations using a paired two-tailed t-test for samples with equal variances. Task duration was measured from the moment that a participant first viewed a scenario until the point when he or she stopped the eye tracker recording. The normality assumption was confirmed with the Shapiro-Wilk test (W = 0.9785, p = 0.1048), and the equal variance assumption was confirmed with Levene's test (F = 0.1920, p = 0.6622). In Figure 25, the boxplot compares the mean task duration between the scenarios where participants could access recommenders and scenarios where participants could not.

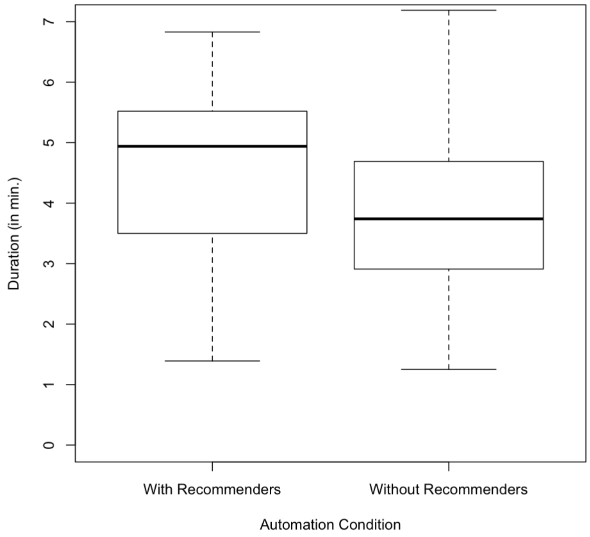

Figure 25. Task duration comparison between experimental conditions

Participants spent slightly less than one minute longer evaluating the display when recommenders were available (μ = 4.563 minutes, σ = 1.33 minutes) than when evaluating the same displays without having access to recommenders (μ = 3.757 minutes, σ = 1.71 minutes). A paired t-test revealed a statistically significant difference between conditions, $t(16) = 4.04$, $p < 0.001$. This suggests that the presence of the recommenders on the display was related to an increase in task duration.

Eye Tracking Metrics

The eye tracking dependent variables (total fixation duration, time to first fixation, and number of fixations) were dependent upon assignment of Areas of Interest (AOIs), or geometric regions surrounding display components that the researcher is interested in evaluating. Using the eye tracking analysis software, Tobii Pro Studio, we created four AOIs, shown in Figure 26. We assessed dependent variables in relation to the core AOIs (MRMS, RP, FFG, and USF).

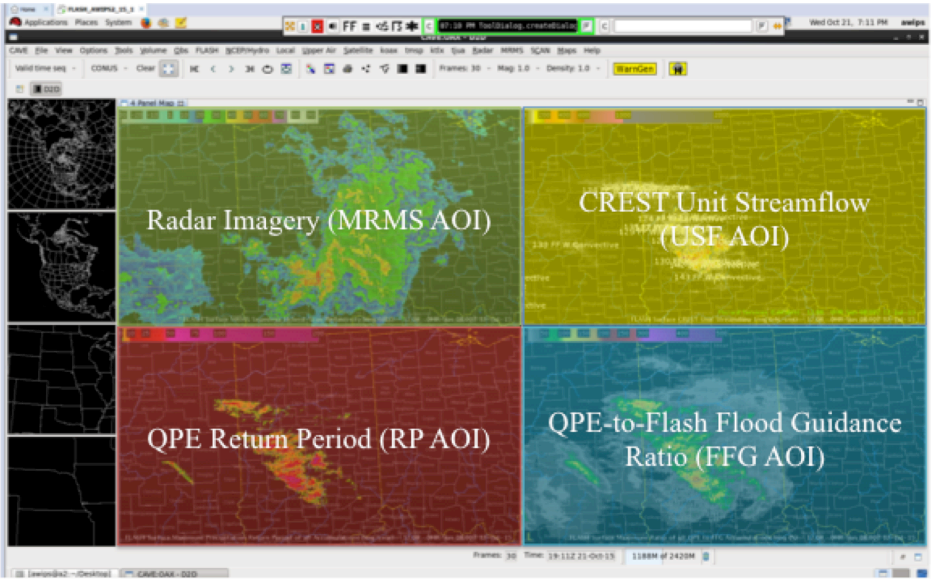
Radar Imagery (MRMS AOI)
CREST Unit Streamflow (USF AOI)
QPE Return Period (RP AOI)
QPE-to-Flash Flood Guidance Ratio (FFG AOI)

Figure 26. AOI assignment used during the analysis

Total Fixation Duration. The total fixation duration measures the amount of time a participant fixated within an AOI over the entire recording period. In this analysis, this measure includes zero values if the participant did not fixate within an AOI. The full dataset satisfied the equal variance assumption (F = 0.5161, p = 0.8222), but it did not satisfy the normality assumption (W = 0.9270, p < 0.001). Outliers beyond 3σ were removed from the full dataset and the tests were run again. The new distribution still failed the Shapiro-Wilk normality test, but maintained equal variance. Thus, differences in total fixation duration were assessed using high breakdown and high efficiency robust linear regression and a Robust Wald test.

Although standard linear regressions frequently use the ordinary least squares (OLS) method for estimating effects, robust linear regression parameters can be based on several different estimation algorithms. Here, we chose MM-estimation with the bisquare weighting function; this technique uses iteratively reweighted least squares

(IRLS) to assign weights to the residuals, is appropriate for nonparametric data, and is also robust against outliers.

After fitting the robust linear regression, shown Table 13, a Wald test based on the robust linear regression coefficients identified a significant main effect of the AOI variable (Wald = 21.72, *p* < 0.001). The main effect of recommender presence was not significant (Wald = 0.5154, *p* = 0.4728) and the interaction between the AOI and recommender condition was not significant (Wald = 0.4175, *p* = 0.5182). In Figure 27, one can see that the mean total fixation duration within the MRMS AOI (the radar data) was greater than fixation duration in any of the other AOIs.

Percentage of Total Fixation Duration. Total fixation duration measured total time spent within an AOI; however, as task duration differed among participants, an alternative measure was needed to normalize fixation patterns. Moore and Gugerty (2010) found that taking the percentage of total fixation duration relative to task duration was a significant predictor of SA. In line with this, we assessed percentage of total fixation duration to determine whether this confirmed the previous conclusions related to total fixation duration.

Table 13. Robust linear regression parameters for the total fixation duration data

	Estimate	**Std. Error**	**t-value**	**Pr(<|t|)**
(Intercept)	36.844	4.055	9.086	< 0.0001
Rec[Available]	4.281	5.963	0.718	0.473
AOI[MRMS]	26.991	5.792	4.66	< 0.0001
AOI[RP]	3.729	5.770	0.646	0.519
AOI[USF]	-4.796	5.778	-0.83	0.407
Rec[Available]:AOI[MRMS]	-8.117	8.508	-0.954	0.341
Rec[Available]:AOI[RP]	-12.572	8.522	-1.475	0.141
Rec[Available]:AOI[USF]	-5.108	7.865	-0.649	0.517

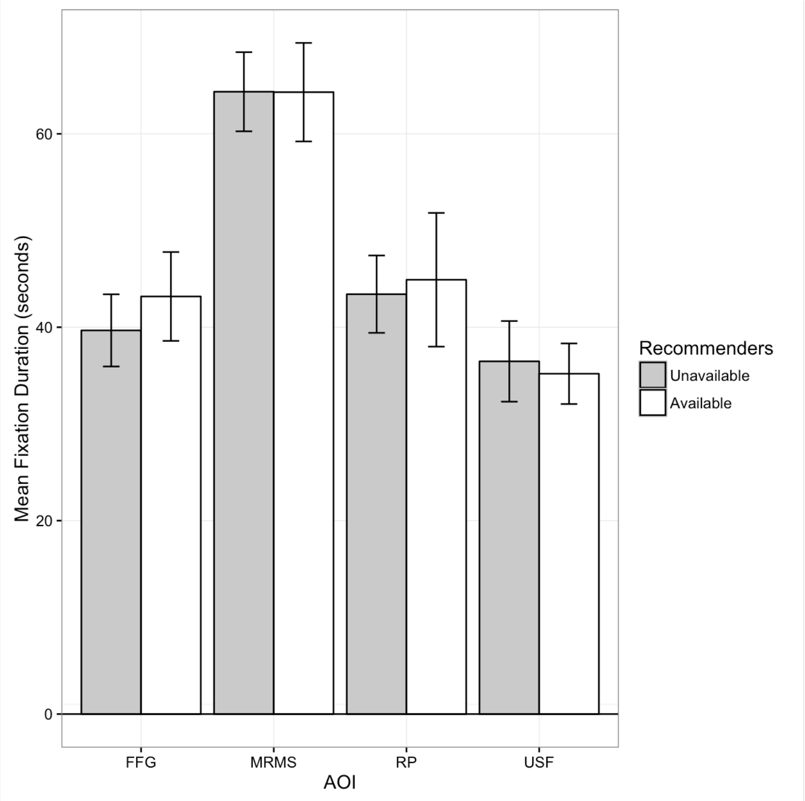

Figure 27. Mean total fixation duration by AOI type

After removing two statistical outliers, the data fulfilled the equal variance assumption (F = 1.902, p = 0.06976) but not the normality assumption (W = 0.9759, p < 0.001). As with the previous analysis, a robust linear regression was used to fit the data. The Robust Wald test failed to detect a significant interaction between recommender condition and AOI (Wald = 0.1889, p = 0.6638), but did identify a significant main effect in proportional fixation duration between AOIs (Wald = 24.23, p < 0.001). However, an alternative approach to robust ANOVA based on a robust F-test did detect a significant interaction between condition and AOI (Robust F = 28.12, p < 0.001). The means displayed in Figure 28, particularly in the USF AOI, lends support to the conclusion that the proportion of time spent fixating across the AOIs changed with recommender availability.

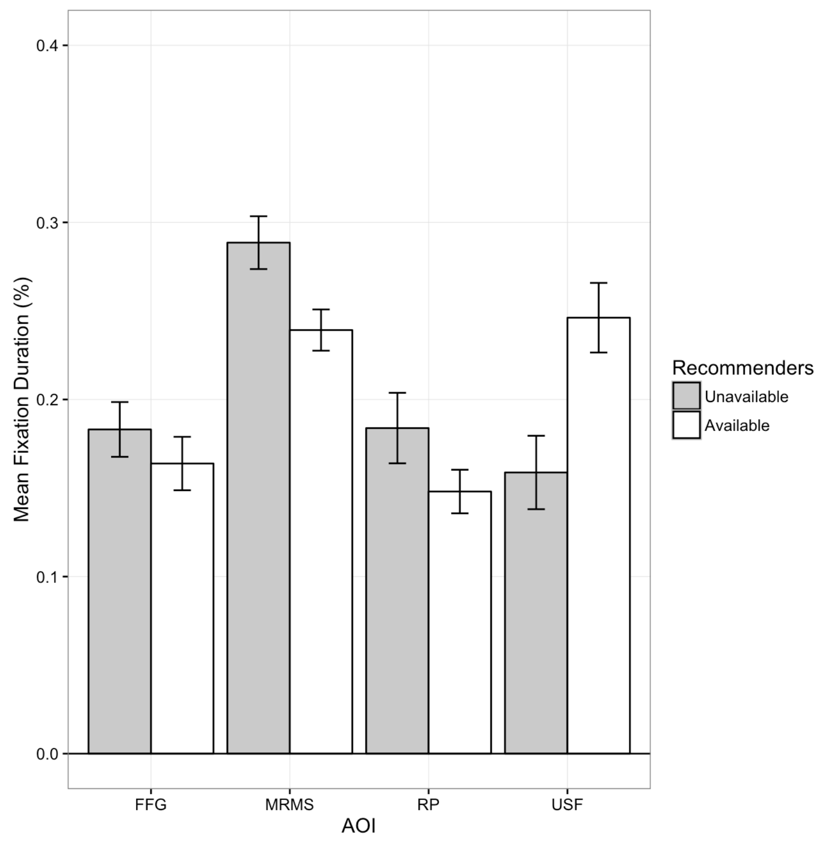

Figure 28. Percentage of fixation duration to task duration by AOI and recommender condition

Figure 28 illustrates that when recommenders were unavailable, participants spent the greatest proportion of their time fixating within the MRMS AOI and the least amount within the USF AOI. However, when recommenders were available within the USF AOI, the proportion of time spent within the AOI increased moderately. In comparison to the total fixation duration results, the normalized results lead to several interesting conclusions. Specifically, the mean absolute time spent within the USF AOI was the lowest in both recommender conditions, yet the proportion of time increased with recommenders.

Mean Number of Fixations by AOI. The number of fixations variable measures the number of times a participant fixated within an AOI. This parameter has been used to indicate the salience or relative importance of an AOI to a decision maker; AOIs with a greater number of fixations may attract a user's attention to a greater degree (Poole and Ball, 2006). Whereas the mean total fixation duration reflects the absolute time that a participant fixated within an AOI, the number of fixations reflects the frequency of fixations within an AOI. The full dataset was not normally distributed (W = 0.8978, $p < 0.001$), but it did satisfy the equal variance assumption (F = 1.791, p = 0.0896). After removing the statistical outliers, the reduced dataset still satisfied the equal variance assumption (F = 1.1817, p = 0.3142) and was normally distributed (W = 0.9881, p = 0.05233).

An ANOVA test indicated that recommender availability produced a significant effect on the number of times a participant fixated during any scenario (F = 4.066, p = 0.045). A significant main effect in the number of fixations between AOIs was also observed (F = 8.031, $p < 0.0001$), although no interaction between the AOI type and recommender condition was found (F = 1.888, p = 0.133); Figure 29 shows the mean number of fixations by AOI type and recommender condition. A Tukey's Honestly Significant Difference test revealed that the number of fixations in the MRMS AOI significantly differed from those in the USF, FFG, and RP AOIs; however, the mean number of fixations within the USF, FFG, and RP AOIs did not significantly differ from each other. This may be due to the visual salience of radar imagery in the MRMS AOI, or alternatively because radar imagery updated every two minutes during the simulations, which was faster than the other AOI types.

Mean Time to First Fixation by AOI. Mean time to first fixation indicates the amount of time it takes a participant to fixate on a particular AOI (in seconds). This can

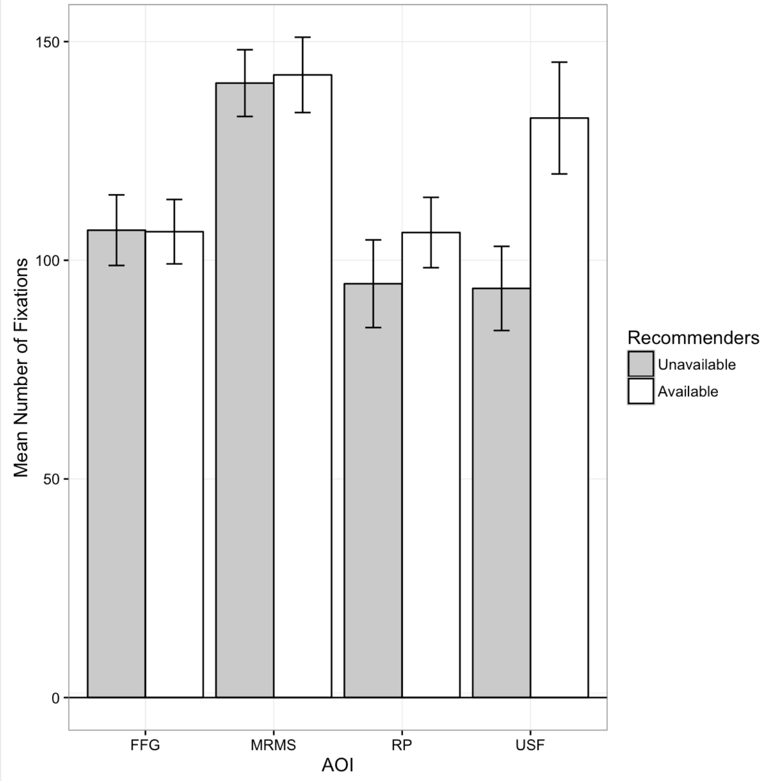

Figure 29. Number of fixations by AOI and condition

be used to interpret the order in which participants viewed AOIs, and could also reflect salience of the information contained with each AOI. Following the removal of statistical outliers beyond 3σ from the mean, the data still did not satisfy the assumption of equal variance (F = 5.600, $p < 0.001$) nor the normality assumption (W = 0.6164*, p* < 0.001). Exponentially transforming the dependent variable led to homoscedastic residuals (F = 0.9673, $p = 0.4554$); however, the residuals were still non-normally distributed, according to the Shapiro-Wilk test (W = 0.1063, $p < 0.001$). Using the transformed data, robust linear regression using MM-estimation and bisquare weighting was used to estimate differences in first fixation time between recommender condition and AOI; the coefficients are shown in Table 14 (Multiple $R^2 = 0.001895$).

Table 14. Robust linear regression parameters for the mean time to the first fixation by AOI and recommender condition

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	525.1	342.2	1.534	0.126
Rec[Available]	-446.7	498.6	-0.896	0.371
AOI[MRMS]	-521.3	449.3	-1.16	0.247
AOI[RP]	-126.2	543.1	-0.232	0.816
AOI[USF]	-284	493.2	-0.576	0.565
Rec[Available]:AOI[MRMS]	561.6	651.9	0.862	0.39
Rec[Available]:AOI[RP]	228.8	761.3	0.301	0.764
Rec[Available]:AOI[USF]	445.2	788.1	0.565	0.573

A Wald test on the robust coefficients failed to identify any significant main effects from AOI (Wald = 1.346, *p* = 0.2459), recommender condition (Wald = 0.8024, *p* = 0.3704), or any significant interaction (Wald = 0.05402, *p* = 0.8162). The mean first fixation time per AOI by recommender condition is shown in Figure 30; from the figure, one can estimate the average order in which AOIs were first viewed. The mean first fixation time on the USF AOI was greatest when recommenders were available. This was an unexpected observation; we anticipated that participants would assess the USF AOI first in this condition, due to the recommender automation’s novelty. Assessing the USF panel after each of other AOIs may suggest that participants were developing SA with the more familiar guidance products, then evaluated the goodness of the automated recommendations with that foreknowledge; this hypothesis, however, is speculative in nature and would require further investigation.

Exploratory Analysis of Scanning Behavior. In addition to recording information related to fixations, the eye tracking system also captured information about gaze direction. Whereas the time to first fixation estimates can reflect the order that participants moved between AOIs, gaze direction analysis reflects scanning patterns

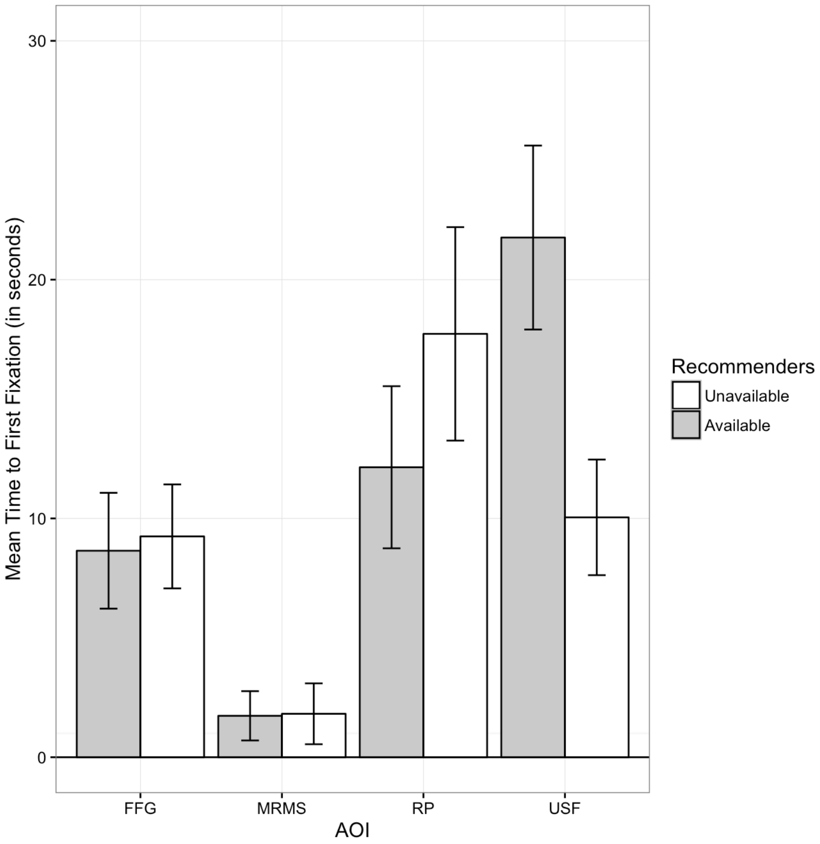

Figure 30. Mean first fixation time versus AOI type and recommender condition among display elements. In the present study, we hypothesized that the recommenders would attract the user's attention to the USF AOI in which the recommenders were embedded in scenarios with the recommender-present condition. We estimated the number of movements between the core AOIs by calculating the frequency of bidirectional exchanges; for example, a fixation within the MRMS AOI followed by a fixation within the USF AOI would count as an MRMS-USF exchange. Exchanges between AOIs and any part of the display not captured in an AOI (e.g. the menu bar at the top of the display) were excluded intentionally.

Figure 31 presents a comparison between gaze movement exchanges between the recommender-present and recommender-absent conditions shown as a percentage of all core AOI exchanges. When recommenders were absent, participants frequently compared the MRMS AOI with the USF and FFG AOIs. When recommenders were present, participants slightly changed their scanning behavior; participants had fewer comparisons between the MRMS and FFG AOIs, but slightly more RP-USF and MRMS-RP exchanges.

Evaluating guidance usage on a participant-by-participant basis revealed individual differences in forecast guidance usage during the forecast decision making process. Figure 32 shows the mean fixation duration percentage by AOI in both automation conditions. While some trends appear, it is clear that each participant had unique assessment strategies. Participants P05 and P13 appeared to have a consistent approach for evaluating guidance products, independent of automation condition. Conversely, Participant P09's assessment approach appeared to be swayed by the

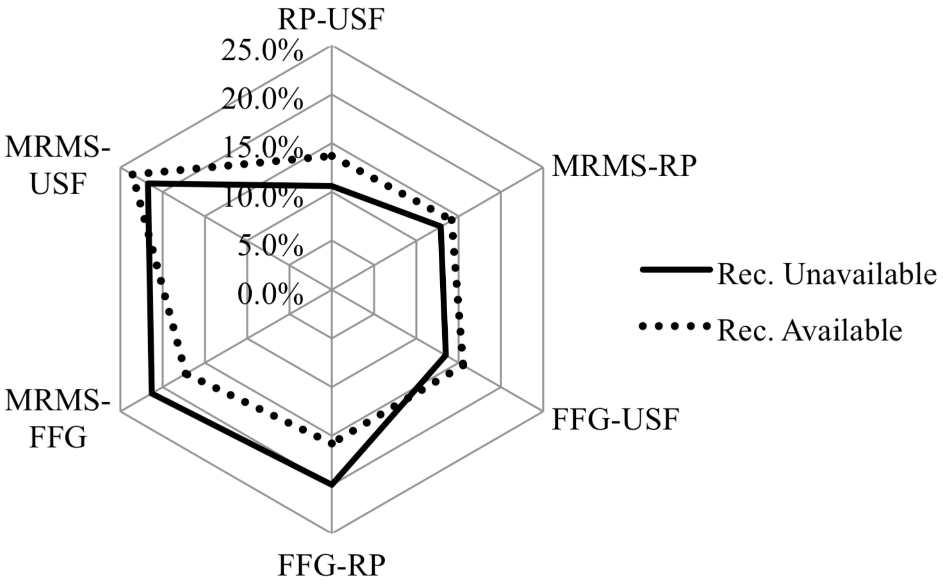

Figure 31. Bidirectional gaze movements between the core AOIs with and without recommenders (always placed in the USF AOI)

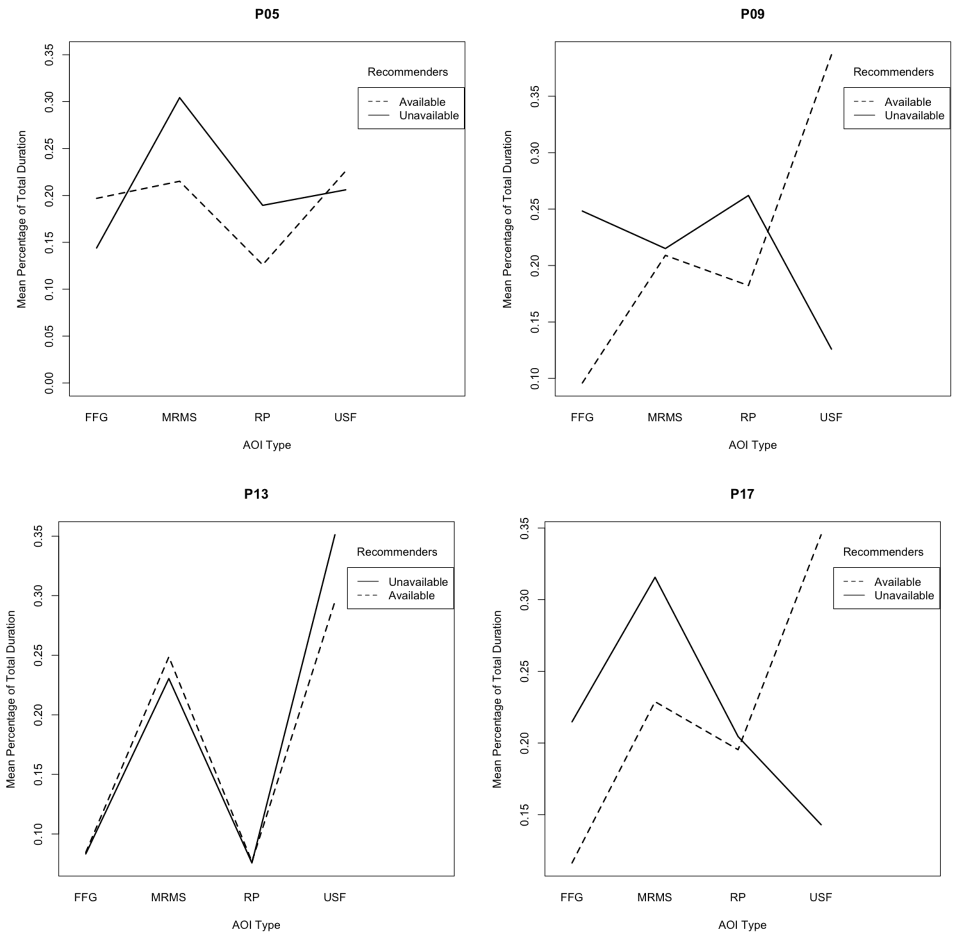

Figure 32. Differences in fixation duration percentage among four participants

availability of recommenders, but she generally relied upon the familiar radar imagery (MRMS AOI). Similar to P09, participant P17 showed interest in the recommenders but not the unit streamflow visualization itself; otherwise, he was fairly consistent in his evaluation strategy. It is possible that some of these differences were due to variations in individual expertise, familiarity with flash flood forecasting guidance products, or level of understanding with regard to the recommenders.

Links Between SA Performance and Eye Movements

Although no differences were observed in SA scores between the recommender conditions, we also hypothesized that eye tracking metrics would predict SA. In order to identify the predictive power of eye tracking variables on the composite SA score, we fit a multiple regression model to the data. The dataset used for the regression was limited in size by the number of eye tracking observations available (n=65). Apart from the eye tracking factors (Total Fixation Duration by AOI (FixDur[AOI]), Percent of Duration by AOI (PerDur[AOI]), Number of Fixations by AOI (n[AOI], and First Fixation Time by AOI (FF[AOI]), we also included the task duration variable (continuous), scenario location variable (categorical, Scenario A, B, and C), and recommender condition variable (categorical, available/unavailable).

Analysis of multicollinearity revealed that six factors (Condition, Scenario, Task Duration, Fixation Duration (USF), Fixation Duration (MRMS), and the Number of Fixations (USF)) and two interaction terms (Condition*Task Duration and Scenario*Task Duration) were highly correlated. As expected, all Total Fixation Duration parameters were highly correlated with the percentage of duration parameters; to overcome this, we fit two distinct models, one using total fixation duration and the other using percentage of duration.

Regression with Total Fixation Duration. We fit the regressions first by fitting the maximal model with all the other non-correlated main effects and two interactions; the results are shown in Table 15. While several coefficients were significant predictors of SA, the Adjusted R^2 value was low (Adj. $R^2 = 0.1703$). Of all the variables included in the maximal model, the only significant main effects identified by an ANOVA were related to scenario location (F = 5.952, *p* = 0.005955) and the Fixation Duration in the RP AOI (F = 7.189, *p* = 0.01111). Thus, the search for a more parsimonious model began.

In order to fit the minimal adequate model, we used both the forward and backward stepwise selection methods. The selection method, in which variables are iteratively added to or removed from the model, terminated at the inclusion of only two variables: the scenario variable and the mean fixation duration within the QPE Return Period (RP) AOI (Adj. $R^2 = 0.4204$). An ANOVA found that both main effects were significant (FixDurRP: F = 29.93, $p < 0.001$; Scenario: F = 9.385, $p = 0.0002843$). The coefficients for the forward selection minimal adequate model are shown in Table 16.

Table 15. Coefficients for the maximal model (predicting composite SA score) after correlated variable removal

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.2730	0.09186	2.972	0.0053
FixDur[FFG]	0.002009	0.002431	0.8260	0.4142
FixDur[RP]	0.001360	0.002379	0.5720	0.5712
n[FFG]	-0.0008008	0.001160	-0.6910	0.4944
n[MRMS]	-0.0004548	0.0005594	-0.8130	0.4218
n[RP]	0.0002446	0.001141	0.2140	0.8315
FF[USF]	0.002134	0.002209	0.9660	0.3406
FF[FFG]	-0.001258	0.001977	-0.6360	0.5288
FF[MRMS]	-0.0006129	0.004405	-0.1390	0.8901
FF[RP]	0.0003372	0.001153	0.2930	0.7716
Rec[Available]	0.003209	0.08228	0.03900	0.9691
Scenario[A]	0.1160	0.08830	1.314	0.1975
Scenario[C]	-0.07205	0.08660	-0.8320	0.4111
Rec[Available]: Scenario[A]	0.02115	0.1173	0.1800	0.8579
Rec[Available]: Scenario[C]	0.06459	0.09879	0.6540	0.5175

Table 16. Regression coefficients for the minimal adequate model (predicting composite SA score) as identified with a forward stepwise selection method.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.2195	0.03680	5.964	< 0.0001
FixDur[RP]	0.002275	0.0005113	4.450	< 0.0001
Scenario[A]	0.1023	0.03414	2.995	0.003980
Scenario[C]	-0.03928	0.03512	-1.118	0.2679

The forward stepwise regression technique selected the RP AOI as a significant predictor, the backward stepwise regression technique selected the MRMS AOI as a significant predictor (at the $\alpha = 0.10$ significance level) of composite SA score. While no effect from automation condition was identified, the results suggest that assessment of specific in-development guidance products can improve overall SA.

Regression with Percentage of Duration. As with the total fixation duration regression, a minimal adequate model was selected with forward and backward stepwise regression method. The forward selection method fitted a model with a higher adjusted R^2 value than the backward selection method. Similar to the previous analysis, the model contained the scenario variable and the percentage of total duration within the QPE Return Period (RP) AOI.

In a secondary analysis to predict SA scores at Level 1, 2, and 3, it was found that the percentage of total duration within the QPE Return Period (RP) AOI was also a significant predictor of SA Level 1 accuracy. Fixation times did not appear to have any significant effect on Level 2 accuracy. However, the SA Level 3 regression suggested that the percentage of total duration within the Unit Streamflow (USF) AOI was a significant predictor.

Confidence Level Analysis

Following data collection, participants completed a debriefing questionnaire. A component of the questionnaire assessed their confidence levels during each scenario. Confidence scores were self-reported on a scale of 1 (strongly disagree) to 5 (strongly agree). Overall, participants reported having a moderate degree of confidence in the Indiana scenario (scenario B; $\mu = 3.89$, $\sigma = 0.66$), the West Virginia/Ohio Valley scenario (scenario C; μ = 3.78, σ = 0.53), and the New Jersey scenario (scenario A; μ = 3.69, σ = 0.77). No significant difference was identified in confidence between locations. As confidence in task performance level was equivalent between geographic locations, this indicates that no scenario was subjectively more challenging than another.

Discussion

This research explored the relationship between SA and recommenders, a type of attention-directing automation. The experiment assessed SA with a probe-based measure, and in addition, we investigated the ability of eye tracking metrics to predict SA in a flash flood forecasting task. Eye tracking is a direct assessment method, yet has had only recent applications in the weather forecasting domain. However, results have shown that the method provides researchers with insight into forecaster information-seeking behavior (Bowden et al., 2016). In the present study, this novel combination of probe-based assessment and eye tracking contributed to a deeper understanding of how graphical attention-directing mechanisms affected forecaster SA.

Independent from SA scores, we hypothesized that participants would fixate on the USF AOI more often and for longer durations when recommenders were available than when they were not. Increases in both measures under the recommender-available condition were expected due to the additional information provided by the recommenders. This hypothesis was partially confirmed; the number of fixations within the USF AOI was significantly greater when recommenders were available than when they were not. This was a logical outcome when one assumes that participants would want to familiarize themselves with the unfamiliar recommenders. However, the fixation duration results also indicated that greater attention to the USF AOI did not noticeably diminish the amount of time participants spent assessing information in the other AOIs. This observation was supported further by the task duration results, which found that when recommenders were available, participants spent, on average, approximately one minute longer assessing the entire dataset than when recommenders were unavailable.

The difference in task duration between conditions may be due to two possible causes. First, the increase in duration may have been related to the additional information presented by the recommenders. Previous research has shown that the forecasting process involves a large amount of comparison between information sources to make sense of the situation and to assess bias (Kirschenbaum, 2004). It is likely that when confronted with the uncertainty surrounding how to use new information sources, the study participants spent the same amount of time reviewing the more familiar AOIs but also spent additional time assessing the recommenders. The evaluation of fixation time percentage lends support to this hypothesis.

The second possibility is that participants took more time to complete the scenarios with recommenders due to the unfamiliarity with the recommenders themselves. During the study, several participants commented that they were not sure that they were using the recommenders correctly. When in such a situation, it is possible that such participants spent more time assessing the information in the recommenders. Still, this supports the expectation that recommenders would draw participant attention from other areas of the display.

While several of the eye tracking and task performance metrics indicated a significant difference in the recommenders-available condition, not all original hypotheses received support. Specifically, the results did not support the hypothesis that the recommenders-available condition would be associated with greater total fixation durations in the USF AOI. Unexpectedly, the USF AOI received the lowest total fixation duration of any of the AOIs whether recommenders were available or not. Participants spent the greatest amount of time fixating within the MRMS AOI (the radar scans). There are several possible explanations for this outcome. This panel held the radar imagery, and it updated more frequently than the hydrologic models did. These updates meant that the visual stimuli changed more frequently, which may have attracted the eyes and motivated participants to reassess the panel more often than the other AOIs. Secondly, the radar imagery had the added benefit of being the most familiar information source available in the study. Participants had varying levels of experience with the other three guidance products.

The comparative nature of situation assessment in weather forecasting has been well established in the decision making literature (Kirschenbaum, 2004; Pliske, Crandall, & Klein, 2004; Trafton & Hoffman, 2007). The gaze movement analysis revealed that some of the most frequent bidirectional exchanges involved the MRMS AOI. Considered alongside the total fixation duration and fixation duration percentage data, this suggests that participants not only used the familiar radar imagery to establish a baseline understanding, but they also used it in comparison tasks, perhaps to calibrate their mental models to the less familiar guidance products. This hypothesis may also be supported by the first fixation time results. On average, participants fixated upon the radar imagery before looking anywhere else on the screen, regardless of the recommender condition.

Implications for Eye Tracking as an SA Assessment Method

The present study suggests that direct, eye tracking-based measures can predict SA accuracy, a finding that is consistent previous research (Moore & Gugerty, 2010; van de Merwe et al., 2012). In a study of air traffic controllers, Moore and Gugerty (2010) found that percentage of time fixating on an AOI was the strongest predictor of overall awareness; the regression analyses in the present study support this. Additionally, Moore and Gugerty (2010) observed that low error rates were associated with higher number of fixations. Based on this precedent, we hypothesized that participants with large fixation durations (both absolute and percentage-based) and fixation counts within the USF AOI would have a higher level of SA than participants with low fixation duration and few fixations with the recommenders.

Similar to the previous studies, we found that the fixation duration percentage was a significant predictor of overall SA. We expanded upon Moore and Gugerty's (2010) work by assessing the predictive power of additional variables and found that the total fixation duration and, to a lesser degree, number of fixations within specific AOIs were alternative predictors. The present findings also weakly support Moore and Gugerty's (2010) observation of the inverse relationship between fixation counts and error rates. The stepwise regression analysis showed that one of the eye tracking measures—number of fixations within the QPE Return Period (RP) AOI—was a significant predictor of overall SA in terms of probe accuracy. This was unexpected, but introduced a new question regarding the utility of other in-development flash flood prediction models. Although radar scans and Flash Flood Guidance-based (FFG) guidance products are available in operational settings, the QPE Return Period (RP) and CREST Unit Streamflow (USF) products are currently in-development and not available for use in formal work display systems. In the experimental scenarios where recommenders were available, the recommender polygons were always overlaid within the USF AOI, so it was intriguing to identify the relationship between frequent use of the QPE Return Period guidance product and SA.

Although this study did not identify a strong relationship between fixation count and SA accuracy, analyses of fixation entropy may provide insight into why participants experienced generally low levels of SA. Entropy, a measure of fixation location variability, has been used to evaluate human attention (van de Merwe et al., 2012). Moore and Gugerty (2010) found that as error rates increased, participants exhibited less focused scanning patterns; the same was found by van de Merwe et al. (2012). Moore and Gugerty (2010) found that successful participants tended to fixate in tight clusters, whereas low-performing participants fixated in seemingly random motions throughout widely spaced areas. High entropy was attributed to uncertain goals as well as high workload (van de Merwe et al., 2012). However, measuring entropy within the weather forecasting display posed a great challenge; whereas aircraft are generally static entities, we question whether this behavior would exhibit itself among weather forecasters.

Air traffic control tasks require focus on distinct areas of interest (e.g. aircraft), but weather forecasting displays often contain many different types of guidance products. Indeed, the four-panel display used in the present study was a simplified version of the type of displays used in operational forecasting; the simplified display was chosen because it allowed for greater control in the experimental design and eye tracking analysis. The decision making literature has suggested that forecasters attend to information in a goal-directed manner (Trafton et al., 2000). Yet, the layout and dynamic nature of many forecast guidance products lead to information attributes changing not only in location, but also in shape, size, and velocity, among others. Thus, we hypothesize that in an operational forecasting task, a weather forecaster's attention would be more dispersed across a display, and the relationship between entropy and error rate may not be generalizable to all domains. Gugerty (2011) suggested that attention allocation capacity could impact an operator's ability to develop SA; future research should explore the role of attention allocation processes on the situation assessment process in weather forecasting.

Implications for Recommender Development

An investigation of the effects of recommender automation on forecaster SA formed the core of this study. We hypothesized that the recommender polygons would act as cues for focal attention, and as such, participants would attend to the highlighted areas and develop more accurate SA than when not exposed to recommenders. While the eye tracking performance metrics did reveal differences in scanning behaviors between automation levels, the probe-based technique did not identify any significant improvements in SA. There are several possible explanations for this outcome, ranging from technological design to individual factors.

In human-computer interaction domain, the literature has suggested that SA and decision making may falter when operators lack experience and trust with technical systems (Kaber & Endsley, 1997). In the present study, participants received an overview of the technical design and purpose of recommenders. In addition, the number of scenarios afforded an opportunity for each participant to work through the initial learning curve. Even so, none of the participants had ever worked with recommenders prior to the study, and several were even unfamiliar with the concept. Discussions of the out-of-the-loop decision making problem have contained cautions against drawing conclusions about the relationship between SA and decision making when operators lack necessary experience (Endsley & Kiris, 1995). Nevertheless, we do not think this solely explains low SA scores in the present work, as participants were all experienced, professional forecasters with some degree of subject matter expertise in hydrologic forecasting.

Apart from a lack of experience, it is possible that the participants did not trust the information provided by the recommenders. In a study of a severe hail recommender system, Karstens et al. (2015) found that forecasters only used recommenders in fewer than 20% of the opportunities in which they were provided, which the authors interpreted as an indication that the forecasters were not interested in using them, or that they didn't trust the recommendations. Studies have shown that low trust in automation can affect operator performance (Hoff & Bashir, 2015; Kaber & Endsley, 1997). Operators can learn to trust a system if it has a transparent and usable design, and if system performance is effective, reliable, and predictable (Hoff & Bashir, 2015). As automated decision support systems for weather forecasting develop and gain more time in use, we believe that user trust in recommender systems would improve with a transparent creation process coupled with a training program.

The recommender system's design may also have influenced participants' abilities to develop SA. Many previous studies have evaluated the impact of level of automation (LOA) on SA; for example, Kaber and Endsley (2004) found that SA level and task performance were best with the aid of low to moderate levels of automation. They hypothesized that SA decrements could have been due to active and passive information processing styles, a conclusion that prefigured Moore and Gugerty's (2010) work. The present findings, in combination with qualitative studies of weather forecasting, suggest that comparison among forecast guidance products and decision support automation involves passive as well as focused processing, but that forecast guidance use may exhibit itself differently in weather forecasting than in other domains, such as air traffic control or driving.

Lastly, it is also possible that the recommenders did affect SA but that the probes lacked the power to detect the differences. Similar to the current work, in an evaluation of LOA and SA, Dao et al. (2009) failed to detect a significant difference in SA probe response accuracy during a short-duration task; however, they observed a significant effect of automation level on response latency measure. A preliminary analysis of response times in completing SA probes in this research, however, failed to detect any significant differences between automation conditions. Following this realization, we hypothesized that the forecast areas in each scenario were too spacious, and that participants may have had high levels of SA for smaller sectors, something not captured in the scoring metric. A case study of the scanning patterns of two low-SA (overall score) participants revealed that they focused their attention on small sections of the forecast area, rarely deviating out to assess other areas of the map. As a result, they received points for the areas of risk they correctly identified, but received no points for the areas of the map they neglected to view. This may have been a mechanism for dealing with the workload, the short timeframe, or it may even have been due to prior professional training. In future work, it would be of benefit to extend this research by assessing SA in sectorized forecast decision making.

Limitations

Several things may limit the generalizability of this study. First, participants were asked to identify areas of risk over regions of the US that they may not have been familiar with. For example, one participant had extensive experience forecasting in Indiana, and so he was very familiar with county names and river structures during the two Indiana scenarios. Others, however, expressed discomfort with being able to assess risk accurately in areas with which they were not familiar. They also had trouble remembering the county names and river structures. As past research into forecasting, experience, and SA has shown, experience is a large factor in SA, and such a lack of experience may be a contributing factor to diminished performance in this experiment.

Technical Limitations. At the time of this study, recommender systems, particularly for flash flooding, were in the early development stages. The recommender algorithm used in the present study was based on a threshold from the QPE Ratio product. The QPE-to-FFG ratio product was chosen due to FFG's familiarity to most forecasters across the United States; however, it might have been better either not to display FFG as one of the four panels in the display, or to select a different recommender product. Even displayed over the unit streamflow (USF) map, it was clear that the recommenders were only highlighting the items that were visible in the

QPE-to-FFG ratio map; however, it did provide a way to directly compare the two products in a way that overlaying the two maps on top of each other did not (overlaying tends to be messy and hard to interpret when the color tables conflict). Forecasters often engage in comparison activities and so this may be a way to facilitate comparison and mental model building (Kirschenbaum, 2004; Trafton & Hoffman, 2007).

In addition, few participants were familiar with the concept of recommender systems, which likely impacted general understanding in terms of usage. In the understanding that recommender systems would become more complex prior to operationalization, we did not want to mislead participants in terms of the current recommender capability. It is possible that participants may not have fully understood how to incorporate recommenders into their decision process. Nevertheless, the eye tracking measures showed that participants did consult the recommender AOI in those scenarios where recommenders were available. As recommender systems take on a higher degree of complexity, their place within the situation assessment process may adapt. Due to technical constraints imposed by the current state of recommender technology, the present work was not able to capture such user behavior, but we suggest that this work has implications on understanding the relationship between SA and graphical mechanisms for directing attention in a weather forecasting display.

Threats to Internal Validity. Several measures were taken in order to reduce threats to the study's internal validity. In terms of a maturation effect, the study was designed to be approximately one hour in duration, and in the longest case, the participant in question took about one and a half hours to complete the study activities. This timeframe was selected to minimize effects from participants' moods or behaviors changing due to tiredness, inattention, or other factors caused by the experiment. Likewise, in order to determine whether the results were affected by testing effects, the researcher conducted a statistical test to identify potential differences in responses caused by order. As previously discussed, no significant difference in terms of fixation duration was found.

Lastly, as with many human-subjects studies, a possible threat from participant reactivity exists. As the experiment occurred in a controlled laboratory environment, and participants were aware that the procedures used eye tracking and an experimental form of automated guidance, it is possible that participants altered their typical situation assessment strategies as a result. In order to combat this, participants were instructed to review the provided data as they usually would. Nevertheless, several participants commented that using the four-panel AWIPS-II display was akin to "trying to work at someone else's desk." In future work, a longitudinal study or a naturalistic decision making approach could be used to assess effects of recommenders with participants in a more natural setting, though this, too, might produce observational bias.

Summary

The current study points to several avenues for future work that could improve understanding about how weather forecasters develop and maintain SA, particularly during tasks involving decision support automation. While the probe-based technique did not reveal a significant effect from the recommender system on overall SA accuracy, the eye tracking analysis did reflect differences in scanning behaviors when participants had access to the automation. From these results, we can conclude that the recommender system did not distract attention from areas of risk; with that in mind, participants largely had low SA scores across all levels. Going forward, additional work should extend the methodology to forecasting over a smaller sector. Furthermore, using eye tracking with a more naturalistic approach with eye tracking may permit researchers to capture data from forecasters in operational settings in a relatively non-intrusive manner. This type of approach could provide insight into real-time situation assessment in scenarios with real-world impacts.

The greatest contribution of this work is the validation that eye tracking can be used to assess SA in weather forecasting, a complex sociotechnical work domain. In the few other studies that have explored the use of eye tracking as an SA assessment method, research has frequently focused on aviation, air traffic control, and driving. The current findings provided direct evidence of the impact of focused attention on forecaster SA, and from a methodological perspective, showed that several eye tracking metrics can be used, to varying degrees, as predictors of SA. This research suggests that eye tracking has potential for use in operational forecasting environments, and it may be an effective tool for assessing training needs for forecasters.

Chapter 6: Discussion and Conclusions

Weather forecasters frequently work with large quantities of data, and previous research has sought to address concerns related to information overload, increased workload, and diminished performance (Daipha, 2010; Karstens et al., 2015; Stuart, Schultz, & Klein, 2007). In the present work, we have been concerned with understanding technological design factors that influence situation awareness (SA) and decision making during flash flood forecasting tasks. Here, we have argued that the forecasting process and outcomes can be improved by designing decision support technology to suit the cognitive and task-related needs of forecasters. In Chapter 3, we investigated the effects of visualization algorithm and visual display properties on perception and recognition of a flash flood threat. In the following chapter, we explored the relationship between information requirements, situation assessment, and uncertainty management along the forecasting timeline. Finally, in Chapter 5, we analyzed the effects of an automated decision support technology on SA with a novel eye tracking methodology.

The current work sought to enhance current understanding of decision making with several studies focused on identifying interactions among decision support design, weather forecasting, and SA. Each study contributed to the user-centered development of the Flooded Locations and Simulated Hydrographs (FLASH) suite of flash flood forecast guidance products. In the first study, we found that data aggregation methods affected signal detection during a flash flood prediction task. Signal detection, a component of Level 1 SA (perception), involves an operator determining whether or not two stimuli are different (Stanislaw & Todorov, 1999). When evaluating guidance products during a weather event, a forecaster must be able to detect patterns and comprehend their significance. Based on the findings, we concluded that the data aggregation method did influence incidence rates of error types in a threat detection task within one of the FLASH visualizations. Participants were statistically most likely to make correct threat assessments when the stimulus event had small spatial coverage, minimal property damage, and was visualized with the average-based aggregation method.

The mixed methods analysis in Chapter 4 led to several insights into information seeking and uncertainty management behaviors during flash flood forecasting. In the quantitative component, we established that forecasters had different SA requirements at different scales, particularly with regard to time and environmental activity level. Like Daipha (2010) and Morss and Ralph (2007), we observed that forecasters compared a diverse collection of guidance products prior to making a forecast decision. Out of all the decision support products available to participants in the Hazardous Weather Testbed experiment, radar imagery, hydrologic models, and flash flood guidance-based guidance products were viewed for the most time prior to issuing a watch, whereas radar imagery, flash flood guidance, and quantitative precipitation estimates (QPE) guidance products received the most screen time in the warning phase. These observations support Kirschenbaum (2004), who suggested that weather forecasters construct their mental models by comparing information sources and extracting information.

In addition to data comparison and goal-directed information extraction, the forecasters attempted to develop their SA under uncertain conditions by employing uncertainty management tactics. Lipshitz and Strauss (1997) proposed that decision makers cope with uncertainty through reduction, assumption-based reasoning, weighing alternatives, forestalling, and suppression, known as the R.A.W.F.S. heuristic. Using a theoretical coding framework based on work by Lipshitz and Strauss (1997), we found that forecasters frequently discussed individual-level tactics, such as reduction and suppression, as well as organizational-level tactics, such as acknowledgement. These findings indicate that when under uncertainty, situation assessment involves more than intuition and analysis, as proposed by others (Doswell, 2004). With examples drawn from the focus group discussions, we illustrated that SA in weather forecasting is in part governed by an individual's ability to cope with uncertainty.

The mixed methods study gathered information about situation assessment and SA requirements for flash flood forecasting in a naturalistic environment, and the third study built upon this foundation to assess the effects of automation on comprehension and projection, while also testing the efficacy of eye tracking as an SA predictor. Whereas the first two studies primarily assessed elements of Level 1 (perception) and Level 2 (comprehension) SA, the final study evaluated SA across all three levels in an automation-aided forecasting task. We hypothesized that availability of flash flood recommenders would draw forecaster attention away from other areas of risk, leading to an operator-out-of-the-loop phenomenon and diminished awareness. This hypothesis was not supported; no statistical differences were observed in SA scores collected with a probe-based technique. This finding suggests that recommenders hold promise as decision support tools for weather forecasting. Although we had hypothesized that recommenders would reduce SA, the results indicated that this did not occur. Still, participants across the board exhibited low levels of SA, which was attributed to large forecasting sectors and levels of local knowledge among participants.

In relation to the literature, these findings are intriguing. While Endsley and Kiris (1995) observed a negative correlation between automation level and SA, Kaber and Endsley (2004) found that moderate levels of automation were associated with higher levels of SA. We suggest that, similar to Kaber and Endsley's (2004) work, the recommenders acted at a moderate level of automation. Eye tracking showed that participants did fixate upon the automation, but as the study did not require participants to base decisions on the automation alone, we propose that SA and task duration would improve as operators become more experienced with the new systems. Particularly in terms of task duration, as duration decreases, forecast lead time has the potential to increase. In addition, while Dao et al. (2009) did not detect any significant difference in SA accuracy between automation levels, they did find that SA as measured by response time did correlate to automation level. A preliminary investigation did not reflect a significant correlation between response times and the current results, but further research would be needed.

Limitations

As discussed in the prior chapters, each study had several limitations. Primarily, sample sizes, participant expertise, and technical issues had the greatest potential to limit the findings and generalizable conclusions. Here, we reflect upon several of the limitations and discuss their relative importance.

Conclusions from the research were constrained by attributes of the samples. In the data aggregation method evaluation (Chapter 3), the sampled participants did not possess professional expertise, which may reduce the ability to draw generalized conclusions about signal detection by expert forecasters. While the participants were not professional forecasters, the sample was selected from the population of meteorology students and postdoctoral researchers at the University of Oklahoma. We justified this choice based on the assumption that those engaged in meteorological studies would have relevant experience with regard to interpreting map-based environmental visualizations.

Participant experience may have also limited performance in the recommender evaluation (Chapter 5). In weather forecasting, an SA requirement is awareness of regional geography, known as local knowledge. While a sample of expert forecasters was used, levels of regional forecasting experience varied. As overall SA scores were generally low, it is possible that some participants lacked geographical knowledge on some of the areas used in the scenarios. Likewise, performance may have also been limited by large forecasting sectors.

In the mixed methods study (Chapter 4), the small sample may cause concern that not all viewpoints were captured in the focus groups. It is worth noting that although the sample size was smaller than the literature has recommended, the qualitative data achieved saturation. There is concern that dominant perspectives can overshadow alternative, minority comments in small sample focus groups. This limitation was considered in the study design and was addressed with proper facilitation techniques. However, in the quantitative analysis of SA requirements, sampling issues may have had a greater impact. Despite having the potential to collect more than 800 hours of forecasting data, issues with the recording software and external influences (e.g. participants using personal computing devices) led to a greatly reduced sample. In order to accommodate for this, data quality was ensured with sampling techniques that removed inadequate recordings. Still, conclusions drawn from this data only reflect a partial image of information seeking behavior during flash flood forecasting.

Upon reflection, these limitations could be overcome through several means in future work. Conducting additional data collection of forecast guidance usage in future testbed studies could add to the data corpus. The same would apply for future applications of the focus group methodology. Circumventing limitations associated with participant experience poses a greater challenge, however. Presenting experienced participants with scenarios located in their regions of employment risks biasing performance if participants recall the actual weather event. As such, we recommend that future studies consider using non-invasive observational techniques or qualitative research methods, such as Cognitive Task Analysis, to assess SA and decision making in an operational or quasi-operational forecasting environment.

Decision Making in Weather Forecasting

Through this work, we have provided a complementary perspective to Doswell's (2004) account of decision making heuristics and biases in weather forecast decision making. Doswell (2004) presented a framework of forecast decision making in which forecasts were determined through a combination of analytical and intuitive processes. Analytic decisions were procedural and rule-based, whereas intuition-based decisions were subject to a number of cognitive biases and heuristics, such as the availability and representativeness heuristics. Supporting this framework, Stuart et al. (2007) argued that a successful integration of analysis and intuition is at the heart of recognition-primed decision making (RPD). Similarly, as a result of interviews with professional forecasters, Pliske, Crandall, and Klein (2004) found that experts were distinguished from non-experts in that they used a more flexible approach to forecasting, seeking and comparing information sources in order to build a model. Non-experts had a more fixed approach, and primarily relied on computational predictions and procedural policy. In the present work, findings from the focus group analysis corresponded to these earlier works and provided an extended understanding of decision support technology's effects on SA under uncertainty.

Situation awareness (SA) is not only regarded as an influential factor in decision making, but much research has focused on its impacts on safety. While some research has focused on effects of SA loss in aviation accidents (Endsley, 1995b; Endsley & Garland, 2000), we argue that the implications from loss of SA in weather forecasting can have just as great impacts. In a review of the warning operations during the May 3, 1999 tornado in Moore, Oklahoma, Andra, Quoetone, and Bunting (2002) attributed forecaster SA to tightly-coupled interactions between decision support technology and individual expertise. The authors associated insufficient SA with conditions involving situational uncertainty and incorrect forecaster preconceptions. Insufficient SA was compounded by information that was "changing and sometimes unexpected, ambiguous or conflicting" (Quoetone, Andra, Bunting, & Jones, 2001). In the current work, observations revealed that forecasters cope with such uncertainty primarily through reduction and suppression tactics, although organizational policy also contributes to uncertainty and risk management. However, this revealed a major challenge related to assessment of SA in the forecasting domain: as framed by several existing SA frameworks, SA is assessed in relation to a "ground truth."

Ground truth has been a point of contention in the literature. Dekker, Hummerdal, and Smith (2010) questioned the appropriateness of accuracy-based assessment techniques, expressing the concern that such a method based would require an "omniscient, normative arbiter or homunculus that knows completely and accurately the interdependencies of all contextually dependent variables" (Dekker et al., 2010). While this perhaps poses slightly less of a problem in the weather domain, where events can be verified objectively after the fact, some observational systems may still contain inaccuracies. This is relevant to the present work, where "ground truth" was determined from historic records and environmental sensor networks (e.g. stream gages), which may contain incomplete entries or imprecise recordings. In our view, minimizing the effects of the role of ground truth on probe-based assessment methods is deserving of attention in future research.

It is possible that a solution could involve reframing "ground truth" in SA assessment to align with definitions of forecast goodness. In domains involving high levels of uncertainty, establishing a situational model mirroring the actual environment may pose a great challenge. For example, in long-term weather predictions, a forecaster may only have access to climatological data and environmental models, and if initial conditions or modeling parameters involve even a small degree of error, predictions can diverge from the eventual actuality. An additional element of complexity is introduced when shifting from deterministic to probabilistic forecasts. One may then ask, is it appropriate to speak of "good SA" if a forecaster's situational model is dependent upon predictions that lack accurate, real-time observations?

In response to this, we suggest that future studies of SA in weather forecasting consider the definition of forecast goodness as discussed by Murphy (1993). In his essay on characterizing forecast quality, Murphy (1993) proposed that a "good" forecast was one that best matched current model predictions, led to societal benefit, and most relevantly, conformed to the forecaster's best conception of the current situation. If one assumes the normativist perspective, then one believes that there is a ground truth for all decisions (Dekker et al., 2010; Parasuraman, Sheridan, & Wickens, 2008). While recognizing that SA, action selection, and performance are governed by different cognitive mechanisms (Wickens, 2015), we question what is the most appropriate type of ground truth to use for assessment. As Endsley (1995c) has pointed out, accurate SA does not necessarily equate to good performance. Indeed, the opposite could easily be imagined, in which a forecaster has a high level of awareness of what the models predict, and then makes a forecast matching their SA, only in retrospect realizing that the prediction was overinflated (or perhaps worse, missed). Based on the present and past research, we suggest that SA is a core component of weather forecast decision making, but accurate assessment of it requires further discussion within the scholarly community.

A Reflection Upon Situation Awareness in Theory and Practice

In addition to exploring situation assessment in weather forecasting, the findings provide insight into human reasoning and judgment under uncertainty. Underpinning this dissertation was Endsley's 1995 Model of SA, which provided a consistent and widely accepted framework for the design of each experiment. In conducting the present work, we identified several aspects of the construct that have as of yet received little attention in the literature.

Several of the longstanding tenets of Endsley's 1995 Model of SA were observed in the present work. In the recommender evaluation, we determined that while forecaster SA was not diminished when exposed to the automation, the decision support tool did influence certain scanning measures. In Endsley's (1995c) model, automation is identified as a limiting factor to SA in many decision environments. However, Kaber and Endsley (1997) also found that when automation allowed interactivity between the operator and the system, SA improved. This difference in levels of automation provides a likely explanation for the present findings.

Additionally, Endsley's 1995 Model of SA recognizes the impact of interface design on an operator's ability to acquire and maintain SA. Indeed, Endsley (1995c) states that interface design affects "how much information can be acquired, how accurately it can be acquired, and to what degree it is compatible with the operator's SA needs." In the data aggregation evaluation in Chapter 3, we found that the algorithm used to aggregate points within a large dataset influences an operator's Level 1 and Level 2 SA. Although the current goal was to identify an aggregation algorithm that minimized missed forecasts, we suggest that these findings could be applied to conform to one's own risk tolerance level.

The present findings also support recent arguments in favor of the nonlinearity of the three levels of SA. Perhaps misconstrued due to the term "level," some scholars have presumed that the Level 1 precedes Level 2, which in turn precedes Level 3 SA

(Chiappe, Strybel, & Vu, 2012; Klein, 2015b). However, in a recent treatise to address misperceptions of the 1995 Model, Endsley (2015) argued that the three levels of SA are not necessarily a linear sequence. While it is easy to view the three levels of SA as a process, our findings support Endsley's (2015b) viewpoint. The present findings support previous work that has showed that situation assessment involves both bottom-up and top-down processing (Endsley, 1995c). The findings, particularly from the mixed methods analysis, suggested that Level 3 SA preceded a goal-directed search for information, which was then coded into Level 1 and Level 2 SA. The present work supports the application of the 1995 Model of SA to the field of forecast decision making, an explanatory framework already accepted by practitioners within the weather community (Jones, Quoetone, Ferree, Magsig, & Bunting, 2003; Quoetone et al., 2001). Nevertheless, this dissertation has shown that several outstanding issues need further investigation.

In the present and related works, the question of external props has consistently stimulated debate. Recent additions to the field, such as the Situated Approach to SA, draw their origins in part from the idea that SA can exist in the environment, not solely within the head, as they suggest the 1995 Model of SA proposed (Chiappe, Strybel, et al., 2012; van Winsen & Dekker, 2015). Addressing misperceptions of the 1995 Model of SA, Endsley (2015b) argued that while SA is affected by situational context, the construct is meaningless if it is not contained within memory. The link between memory and SA has been well established in the literature (Endsley, 1995c; Adams, Tenney, and Pew, 1995; Wickens, 2015) but we suggest that the role of props needs examination. For example, in the present work, the freeze-probe measure revealed low levels of SA among the participants. We hypothesized that this may have been due to a combination of large forecasting sectors and a divergence between reality and the experimental design. Specifically, we observed participants having difficulty orienting to the blank maps presented in the probes. It is possible that had participants been given access to the displays while answering the probes, performance would have improved. Indeed, this would be a more familiar process for professional forecasters; few, if any, operational forecast decisions without an external prop (e.g. a radar scan or a numerical model output) visible on the display. Thus, while SA acquired from additional guidance products would theoretically be contained within the forecaster's memory, the awareness and subsequent action selection does not occur independent of the environmental props and cues. We concur with Endsley (2015a, 2015b) that inanimate displays do not possess SA, but we do suggest that further discussion is needed in order to establish a role for memory props in existing frameworks.

Interestingly, the systems-level perspective on SA (also known as the Joint-Cognitive Systems approach) may provide some insight into the question of props and cues. The scope of this dissertation was delimited to the level of the individual decision maker. However, operational forecasting environments involve a high degree of interaction among actors and technological systems. Although not discussed at length within this work, further investigation of how SA is distributed throughout the weather forecasting system would contribute to a greater of understanding of the factors that influence how SA is dynamically distributed throughout the weather domain. The Team SA framework corresponds to the 1995 Model of SA, and as such, has received considerable attention within the literature related to group decision making and performance (Endsley & Jones, 2001; Kaber & Endsley, 1998; Salas, Prince, Baker, & Shrestha, 1995). While yet to receive the level of attention that has been given to the 1995 Model of SA, newcomers to the field, such as Distributed Situation Awareness (Salmon, Stanton, Walker, Jenkins, and Rafferty (2010); Stanton et al. (2006)) and the Situated Approach to SA (Chiappe, Rorie, Morgan, and Vu (2012); Chiappe, Strybel, et al. (2012)) offer new insights into SA at different levels of decision making. The foundation for both these frameworks, the Joint-Cognitive Systems perspective frames SA as an emergent property of a system in which the human operator is a part. We do not suggest that research should only be concerned with decision making at the individual- or the systems-level; simply, we argue that in order to gain a more comprehensive picture of the factors and processes that influence SA, both levels require study.

Future Work

The exploration into SA in weather forecasting revealed several new questions that could extend models of situation awareness in complex systems. Particularly in the automation study (Chapter 5), we hypothesized that the resulting low SA scores may have been due, in part, to the probe technique: a modified Situation Assessment Global Assessment Technique (SAGAT; Endsley (1995a)). When designing the experiment, we selected a freeze-probe technique in order to test the assumption that SA was construct limited by individual memory. In Endsley's (2015b) view, "information that exists in the environment… but of which the operator is not aware… does not constitute SA. It is by definition information of which he or she is not aware (hence the opposite of SA)." However, in the automation study as well as the observational work described in Chapter 4, we observed that forecasters appeared to rely on external cues within the decision support systems during the reasoning and judgment processes. Therefore, we recommend that future research should extend the method presented here with an alternative freeze-probe technique, such as the Situation Present Assessment Method (SPAM), which allows operators to access relevant data in order to respond to probes.

In addition to contributing to a more comprehensive understanding of individual cognition, research should also examine the sociotechnical and interactive aspects of situation awareness. Although not the focus of the present work, weather forecasting involves a large degree of interaction between human operators and technical systems. Particularly, the theoretical frameworks provided by Team SA (Endsley & Jones, 2001; Kaber & Endsley, 1998) and Distributed Situation Awareness (Stanton, Salmon, Walker, & Jenkins, 2009; Stanton et al., 2006) may partially describe factors affecting the transmittance of SA among decision makers within the forecasting system. Viewing SA as a systems-level construct, as does the Distributed Situation Awareness (DSA) theory, may have bearing on the study of weather forecasting. We suggest that a comparative evaluation of theoretical models of SA, including DSA, Team SA, and Endsley's 1995 Model, could build on the present findings and ultimately lead to new insights into SA and reasoning within the weather forecasting domain. Such an evaluation could provide guidance for training, best practices, and policy in the operational forecasting environment.

From a methodological perspective, the present work confirmed previous research regarding eye tracking as a direct measure of SA. Like Moore and Gugerty (2010), we found that fixation duration predicted SA accuracy. While eye tracking allowed for direct inspection of situation assessment and information scanning behavior, it is possible that constraints imposed by the laboratory environment affected experimental outcomes, to some degree. We suggest that incorporating an eye tracking method, similar to that used in Chapter 5, within a study framed with the naturalistic decision making philosophy would theoretically overcome potential experimenter bias while allowing forecasters to work with decision support systems with which they have experience.

Still, limitations within the weather forecasting and environmental sensing system add further challenges to adequate assessment of SA. Scholars have previously pointed out that establishing a ground truth for every situation can be difficult, even unfeasible, in some situations (Dekker et al., 2010). While establishing the "ground truth" in weather forecasting is possible, limited observational sensor systems and verification challenges constrain current ability to receive feedback rapidly during forecasting (Gourley et al., 2013). Indeed, as some of the present focus group findings indicated, verification of a flash flood may not even occur if the potentially affected area was remote and unlikely to cause direct impact to humans or property. Furthermore, in weather forecasting tasks, "ground truth" may not even be possible to perceive accurately with existing technology; as has been cautioned throughout this work, decision support tools—particularly computational models—involve varying degrees of uncertainty introduced by incomplete understanding of environmental processes and uncertain initial conditions (Doswell, 2004). In weather forecasting, developing an accurate situational model involves continuous comparison and assessment of information sources. It is our view that until environmental observational systems mirror those found in systems like air traffic control or aviation, forecaster SA will be limited to what can be gathered from an incomplete representation of the environment.

Conclusion

From a theoretical perspective, this work contributed to a greater understanding of situation awareness in complex systems involving uncertainty. Furthermore, this research was able to contribute to the development of a flash flood prediction decision support system through its transition from research to operations. While some have suggested that it is more appropriate to view SA at a systems-level when studying collaborative, sociotechnical systems (Salmon et al., 2008; Stanton, Salmon, & Walker, 2015), we argue that for weather forecasters, understanding SA at an individual level can have several benefits. From a practice-oriented perspective, studying the relationship between SA and technology may provide foundations for design improvements to decision support systems (Endsley & Hoffman, 2002). Secondly, designing systems to support situation assessment has been linked to performance, specifically in terms of workload- or attention-related errors (Klein, 2000).

Based on current findings, we conclude that weather forecasting decision support systems assist operators in coping with uncertainty in order to acquire and maintain situation awareness. Decision makers leverage technology to maintain SA, but it is also important to minimize overreliance upon such systems, which can also lead to errors. This work described SA and situation assessment in the weather forecasting domain, but findings and recommendations may provide insight in additional decision making environments where uncertainty is high, such as emergency medicine or military command and control. To expand the power of the present findings, we recommend that future research should attend to the interactive nature of situation assessment in the weather forecasting domain. We recommend that development of decision support systems, particularly those for weather forecasting, should incorporate a user-centered design phase. SA errors and performance limitations may be reduced with increased attention to the effects of interface design and visualization methods on human decision making.

Works Cited

Adams, M. J., Tenney, Y. J., & Pew, R. W. (1995). Situation Awareness and the Cognitive Management of Complex-Systems. *Human Factors, 37*(1), 85-104. doi:Doi 10.1518/001872095779049462

American Meteorological Society. (Ed.) (2015) Glossary of Meteorology. Retrieved from http://glossary.ametsoc.org/wiki/Precipitable_water.

Andra, D. L., Quoetone, E. M., & Bunting, W. F. (2002). Warning Decision Making: The Relative Roles of Conceptual Models, Technology, Strategy, and Forecaster Expertise on 3 May 1999. *Weather and Forecasting, 17*(3), 559-566. doi:10.1175/1520-0434(2002)017<0559:WDMTRR>2.0.CO;2

Andrienko, N., & Andrienko, G. (2006). *Exploratory analysis of spatial and temporal data: a systematic approach*: Springer Science & Business Media.

Argyle, E. M., Ling, C., & Gourley, J. J. (2015). Evaluation of Data Display Methods in a Flash Flood Prediction Tool. In S. Yamamoto (Ed.), *Human Interface and the Management of Information. Information and Knowledge Design* (Vol. 9172, pp. 15-22): Springer International Publishing.

Artman, H. (2000). Team situation assessment and information distribution. *Ergonomics, 43*(8), 1111-1128. doi:10.1080/00140130050084905

Barthold, F. E., Workoff, T. E., Cosgrove, B. A., Gourley, J. J., Novak, D. R., & Mahoney, K. M. (2015). Improving Flash Flood Forecasts: The HMT-WPC Flash Flood and Intense Rainfall Experiment. *Bulletin of the American Meteorological Society* (2015).

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Baxter, G. D., & Bass, E. J. (1998, 22-25 Mar 1998). *Human error revisited: some lessons for situation awareness.* Paper presented at the Human Interaction with Complex Systems, 1998. Proceedings.

Belzile, J. A., & Öberg, G. (2012). Where to begin? Grappling with how to use participant interaction in focus group design. *Qualitative Research, 12*(4), 459-472. doi:10.1177/1468794111433089

Beven, K., Cloke, H., Pappenberger, F., Lamb, R., & Hunter, N. (2015). Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface. *Science China Earth Sciences, 58*(1), 25-35.

Bolstad, C. A., Riley, J. M., Jones, D. G., & Endsley, M. R. (2002). *Using goal directed task analysis with Army brigade officer teams.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Bowden, K. A., & Heinselman, P. L. (2016). A Qualitative Analysis of NWS Forecasters' use of Phased Array Radar Data during Severe Hail and Wind Events. *Weather and Forecasting, 31*, 43-55. doi:10.1175/WAF-D-15-0089.1

Bowden, K. A., Heinselman, P. L., & Kang, Z. (2016). Exploring Applications of Eye-Tracking in Operational Meteorology Research. *Bulletin of the American Meteorological Society*. doi:10.1175/BAMS-D-15-00148.1

Bowden, K. A., Heinselman, P. L., Kingfield, D. M., & Thomas, R. P. (2015). Impacts of Phased-Array Radar Data on Forecaster Performance during Severe Hail and Wind Events. *Weather and Forecasting, 30*(2), 389-404. doi:10.1175/WAF-D-14-00101.1

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101.

Braun, V., Clarke, V., & Terry, G. (2014). Thematic Analysis. In P. Rohleder & A. C. Lyons (Eds.), *Qualitative Research in Clinical and Health Psychology*: Palgrave Macmillan.

Braunhofer, M., Elahi, M., Ricci, F., & Schievenin, T. (2013). Context-Aware Points of Interest Suggestion with Dynamic Weather Data Management. In Z. Xiang & I. Tussyadiah (Eds.), *Information and Communication Technologies in Tourism 2014: Proceedings of the International Conference in Dublin, Ireland, January 21-24, 2014* (pp. 87-100). Cham: Springer International Publishing.

Bui, L. (2014, June 10). Heavy rain in Prince George's causes flash flooding; two dozen homes evacuated. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/crime/heavy-rain-in-prince-georges-causes-flash-flooding-several-cars-stalled-in-high-waters/2014/06/10/97857d92-f0ac-11e3-bf76-447a5df6411f_story.html

Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction, 12*(4), 331-370. doi:10.1023/a:1021240730564

Burns, C. M., Skraaning, G., Jamieson, G. A., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of Ecological Interface Design for Nuclear Process Control: Situation Awareness Effects. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(4), 663-679. doi:10.1518/001872008x312305

Bustamante, E. A., Fallon, C. K., Bliss, J. P., Bailey, W. R., III, & Anderson, B. L. (2005). Pilots' Workload, Situation Awareness, and Trust During Weather Events as a Function of Time Pressure, Role Assignment, Pilots' Rank, Weather Display, and Weather System. *International Journal of Applied Aviation Studies, 5*(2), 347-367.

Byrne, E. (2015). Commentary on Endsley's "Situation Awareness Misconceptions and Misunderstandings". *Journal of Cognitive Engineering and Decision Making, 9*(1), 84-86. doi:10.1177/1555343414554703

Caplan, S. (1990). Using focus group methodology for ergonomic design. *Ergonomics, 33*(5), 527-533. doi:10.1080/00140139008927160

Catherwood, D., Edgar, G. K., Nikolla, D., Alford, C., Brookes, D., Baker, S., & White, S. (2014). Mapping Brain Activity During Loss of Situation Awareness: An EEG Investigation of a Basis for Top-Down Influence on Perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 56*(8), 1428-1452. doi:10.1177/0018720814537070

Chiappe, D. L., Rorie, R. C., Morgan, C. A., & Vu, K.-P. L. (2012). A situated approach to the acquisition of shared SA in team contexts. *Theoretical Issues in Ergonomics Science, 15*(1), 69-87. doi:10.1080/1463922X.2012.696739

Chiappe, D. L., Strybel, T. Z., & Vu, K.-P. L. (2012). Mechanisms for the acquisition of situation awareness in situated agents. *Theoretical Issues in Ergonomics Science, 13*(6), 625-647.

Chiappe, D. L., Strybel, T. Z., & Vu, K.-P. L. (2015). A Situated Approach to the Understanding of Dynamic Situations. *Journal of Cognitive Engineering and Decision Making, 9*(1), 33-43. doi:10.1177/1555343414559053

Clark, A. J., Weiss, S. J., Kain, J. S., Jirak, I. L., Coniglio, M., Melick, C. J., . . . Correia, J. (2011). An Overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bulletin of the American Meteorological Society, 93*(1), 55-74. doi:10.1175/BAMS-D-11-00040.1

Clark, R. A., Gourley, J. J., Flamig, Z. L., Hong, Y., & Clark, E. (2014). CONUS-Wide Evaluation of National Weather Service Flash Flood Guidance Products. *Weather and Forecasting, 29*(2), 377-392. doi: 10.1175/waf-d-12-00124.1

Daipha, P. (2010). Visual perception at work: Lessons from the world of meteorology. *Poetics, 38*(2), 151-165.

Daipha, P. (2015). From Bricolage to Collage: The Making of Decisions at a Weather Forecast Office. *Sociological Forum, 30*(3), 787-808. doi:10.1111/socf.12192

Dao, A.-Q. V., Brandt, S. L., Battiste, V., Vu, K.-P. L., Strybel, T., & Johnson, W. W. (2009). The impact of automation assisted aircraft separation on situation awareness *Human Interface and the Management of Information. Information and Interaction* (pp. 738-747): Springer.

Dekker, S. W. A. (2000). Crew situation awareness in high-tech settings: tactics for research into an ill-defined phenomenon. *Transportation Human Factors, 2*(1), 49-62.

Dekker, S. W. A., Hummerdal, D. H., & Smith, K. (2010). Situation awareness: some remaining questions. *Theoretical Issues in Ergonomics Science, 11*(1-2), 131-135.

Dobson, M. W. (1979). Visual information processing during cartographic communication. *The Cartographic Journal, 16*(1), 14-20.

Doswell, C. A. (2004). Weather forecasting by humans-Heuristics and decision making. *Weather and Forecasting, 19*(6), 1115-1126.

Durso, F. T., & Gronlund, S. D. (1999). Situation awareness. *Handbook of applied cognition*, 283-314.

Durso, F. T., Rawson, K. A., & Girotto, S. (2007). Comprehension and situation awareness. *Handbook of applied cognition, 2*, 163-193.

Durso, F. T., Truitt, T. R., Hackworth, C. A., Crutchfield, J. M., Nikolic, D., Moertl, P. M., . . . Manning, C. A. (1995). Expertise and chess: A pilot study comparing situation awareness methodologies. *Experimental analysis and measurement of situation awareness*, 295-303.

Elmqvist, N., & Fekete, J. D. (2010). Hierarchical aggregation for information visualization: overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics, 16*(3), 439-454. doi:10.1109/TVCG.2009.84

Endsley, M. R. (1988a). *Design and evaluation for situation awareness enhancement.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Endsley, M. R. (1988b). *Situation awareness global assessment technique (SAGAT).* Paper presented at the Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National.

Endsley, M. R. (1994). Situation Awareness in Dynamic Human Decision Making: Measurement. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems: Proceedings of a CAHFA Conference* (pp. 79-97).

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 65-84.

Endsley, M. R. (1995b). A taxonomy of situation awareness errors. *Human factors in aviation operations, 3*(2), 287-292.

Endsley, M. R. (1995c). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64.

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 3-32.

Endsley, M. R. (2001). *Designing for situation awareness in complex systems.* Paper presented at the Proceedings of the Second International Workshop on symbiosis of humans, artifacts and environment.

Endsley, M. R. (1997). The role of situation awareness in naturalistic decision making. In C. E. Z. G. Klein (Ed.), *Naturalistic decision making* (pp. 269-283). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Endsley, M. R. (2015a). Final Reflections: Situation Awareness Models and Measures. *Journal of Cognitive Engineering and Decision Making, 9*(1), 101-111. doi:10.1177/1555343415573911

Endsley, M. R. (2015b). Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making, 9*(1), 4-32. doi:10.1177/1555343415572631

Endsley, M. R., & Garland, D. J. (2000). *Pilot situation awareness training in general aviation.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Endsley, M. R., & Hoffman, R. R. (2002). The Sacagawea Principle. *Intelligent Systems, IEEE, 17*(6), 80-85. doi:10.1109/MIS.2002.1134367

Endsley, M. R., & Jones, D. G. (2001). A model of inter- and intra-team situation awareness: Implications for design, training and measurement. In M. McNeese, E. Salas, & M. R. Endsley (Eds.), *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments*. Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R., & Jones, W. (2013). Situation awareness. *The Oxford Handbook of Cognitive Engineering*, 88-108.

Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(2), 381-394. doi:10.1518/001872095779064555

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998a). *A comparative analysis of SAGAT and SART for evaluations of situation awareness.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998b). A Comparative Analysis of SAGAT and SART for Evaluations of Situation Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 42*(1), 82-86. doi: 10.1177/154193129804200119

Evans, C., Van Dyke, D. F., & Lericos, T. (2014). How Do Forecasters Utilize Output from a Convection-Permitting Ensemble Forecast System? Case Study of a High-Impact Precipitation Event. *Weather and Forecasting, 29*(2), 466-486. doi: 10.1175/WAF-D-13-00064.1

Flach, J. M. (2015). Situation Awareness: Context Matters! A Commentary on Endsley. *Journal of Cognitive Engineering and Decision Making, 9*(1), 59-72. doi:10.1177/1555343414561087

Freeman, T. (2006). 'Best practice' in focus group research: making sense of different views. *Journal of advanced nursing, 56*(5), 491-497. doi:10.1111/j.1365-2648.2006.04043.x

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327-358. doi: 10.1037/h0061470

Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P.-E., Clark III, R., Argyle, E. M., . . . Howard, K. (2016). The Flooded Locations And Simulated Hydrographs (FLASH) project: improving the tools for flash flood monitoring and prediction across the United States. *Manuscript submitted for publication*.

Gourley, J. J., Hong, Y., Flamig, Z. L., Arthur, A., Clark, R., Calianno, M., . . . Krajewski, W. F. (2013). A Unified Flash Flood Database across the United States. *Bulletin of the American Meteorological Society, 94*(6), 799-805. doi:Doi 10.1175/Bams-D-12-00198.1

Guerrero, H., Myers, L., Lyons, S., Dunn, J., & Johnson, M. (2015). *Public Reaction to National Weather Service Impact Based Warnings and The Effectiveness of Decision Support Services Provided During the June 12, 2014, Abilene, Texas Extreme Wind and Hail Event*. Paper presented at the 95th Annual Meeting of the American Meteorological Society, Phoenix, AZ.

Gugerty, L. (2011). Situation awareness in driving. In D. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology.* Boca Raton, FL: CRC Press.

Halverson, J. B. (2014, June 11). The June 10 College Park flash flood: How did it happen and why wasn't it forecast? *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/capital-weather-gang/wp/2014/06/11/the-june-10-college-park-flash-flood-how-did-it-happen-and-why-wasnt-it-forecast/

Hegarty, M. (2011). The cognitive science of visual-spatial displays: implications for design. *Top Cogn Sci, 3*(3), 446-474. doi:10.1111/j.1756-8765.2011.01150.x

Hegarty, M., Smallman, H. S., & Stull, A. T. (2012). Choosing and Using Geospatial Displays: Effects of Design on Performance and Metacognition. *Journal of Experimental Psychology-Applied, 18*(1), 1-17. doi:Doi 10.1037/A0026625

Heideman, K. F., Stewart, T. R., Moninger, W. R., & Reagan-Cirincione, P. (1993). The Weather Information and Skill Experiment (WISE): The Effect of Varying Levels of Information on Forecast Skill. *Weather and Forecasting, 8*(1), 25-36. doi:10.1175/1520-0434(1993)008<0025:TWIASE>2.0.CO;2

Heinselman, P. L., LaDue, D. S., & Lazrus, H. (2012). Exploring Impacts of Rapid-Scan Radar Data on NWS Warning Decisions. *Weather and Forecasting, 27*(4), 1031-1044. doi:Doi 10.1175/Waf-D-11-00145.1

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(3), 407-434. doi:10.1177/0018720814547570

Hoffman, R. R. (2015). Origins of Situation Awareness: Cautionary Tales From the History of Concepts of Attention. *Journal of Cognitive Engineering and Decision Making, 9*(1), 73-83. doi:10.1177/1555343414568116

Hoffman, R. R., & Coffey, J. W. (2004). *Weather forecasting and the principles of complex cognitive systems.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Hoffman, R. R., Crandall, B., & Shadbolt, N. (1998). Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 40*(2), 254-276.

Hoffman, R. R., Detweiler, M., Conway, J. A., & Lipton, K. (1993). Some Considerations in Using Color in Meteorological Displays. *Weather and Forecasting, 8*(4), 505-518. doi: Doi 10.1175/1520-0434(1993)008<0505:Sciuci>2.0.Co;2

Hoffman, R. R., Trafton, J. G., & Roebber, P. (2006). Minding the weather: How expert forecasters think: MIT Press Cambridge MA.

Hoffman, R. R., & Woods, D. (2005). Toward a theory of complex and cognitive systems. *Intelligent Systems, IEEE, 20*(1), 76-79.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*(1), 1-55. doi: http://dx.doi.org/10.1016/0010-0285(92)90002-J

Hornbæk, K., & Hertzum, M. (2011). The notion of overview in information visualization. *International Journal of Human-Computer Studies, 69*(7–8), 509-525. doi:http://dx.doi.org/10.1016/j.ijhcs.2011.02.007

Jensen, E. (2009). Sensemaking in military planning: a methodological study of command teams. *Cognition, Technology & Work, 11*(2), 103-118. doi: 10.1007/s10111-007-0084-x

Jones, D. G. (2015). A Practical Perspective on the Utility of Situation Awareness. *Journal of Cognitive Engineering and Decision Making, 9*(1), 98-100. doi:10.1177/1555343414554804

Jones, D. G., & Endsley, M. R. (2004). Use of real-time probes for measuring situation awareness. *The International Journal of Aviation Psychology, 14*(4), 343-367.

Jones, D. G., Quoetone, E. M., Ferree, J. T., Magsig, M. A., & Bunting, W. F. (2003). An Initial Investigation into the Cognitive Processes Underlying Mental Projection. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 47*(3), 596-600. doi:10.1177/154193120304700372

Kaber, D. B., & Endsley, M. R. (1997). Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. *Process Safety Progress, 16*(3), 126-131. doi:10.1002/prs.680160304

Kaber, D. B., & Endsley, M. R. (1998). Team situation awareness for process control safety and performance. *Process Safety Progress, 17*(1), 43-48. doi:10.1002/prs.680170110

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science, 5*(2), 113-153. doi:10.1080/1463922021000054335

Karstens, C. D., Stumpf, G., Ling, C., Hua, L., Kingfield, D., Smith, T. M., . . . Rothfusz, L. P. (2015). Evaluation of a Probabilistic Forecasting Methodology for Severe Convective Weather in the 2014 Hazardous Weather Testbed. *Weather and Forecasting*. doi:10.1175/WAF-D-14-00163.1

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review, 95*(2), 163.

Kirschenbaum, S. S. (2004). *The Role of Comparison in Weather Forecasting: Evidence from two Hemispheres!* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Kitzinger, J. (1994). The methodology of focus groups: the importance of interaction between research participants. *Sociology of health and illness, 16*(1), 103-121.

Klein, G. A. (1989). *Strategies of decision making*. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a226146.pdf

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138-147). Westport, CT, US: Ablex Publishing.

Klein, G. A. (1999). *Sources of power: How people make decisions*: MIT press.

Klein, G. A. (2000). Analysis of Situation Awareness from Critical Incident Reports. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement*: CRC Press.

Klein, G. A. (2008). Naturalistic Decision Making. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 456-460.

Klein, G. A. (2015a). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition*. doi:http://dx.doi.org/10.1016/j.jarmac.2015.07.001

Klein, G. A. (2015b). Whose Fallacies? *Journal of Cognitive Engineering and Decision Making, 9*(1), 55-58. doi:10.1177/1555343414551827

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). *Rapid decision making on the fire ground.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *Systems, Man and Cybernetics, IEEE Transactions on, 19*(3), 462-472. doi:10.1109/21.31053

Klein, G. A., Moon, B. M., & Hoffman, R. R. (2006a). Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE intelligent systems, 21*(4), 70-73.

Klein, G. A., Moon, B. M., & Hoffman, R. R. (2006b). Making Sense of Sensemaking 2: A Macrocognitive Model. *Intelligent Systems, IEEE, 21*(5), 88-92. doi:10.1109/MIS.2006.100

Klein, G. A., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). *A data-frame theory of sensemaking.* Paper presented at the Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making.

Klein, G. A., Pliske, R., Crandall, B., & Woods, D. D. (2005). Problem detection. *Cognition, Technology & Work, 7*(1), 14-28. doi: 10.1007/s10111-004-0166-y

Klein, G. A., Ross, K. G., Moon, B. M., Klein, D. E., Hoffman, R. R., & Hollnagel, E. (2003). Macrocognition. *Intelligent Systems, IEEE, 18*(3), 81-85.

Kreuger, R. A., & Casey, M. A. (1994). *Focus groups: a practical guide for applied research*. Thousand Oaks, CA, USA: Sage.

LaDue, D. S. (2011). *How meteorologists learn to forecast the weather: Social dimensions of complex learning.* (3482746 Ph.D.), The University of Oklahoma, Ann Arbor. Retrieved from http://search.proquest.com/docview/910539812?accountid=12964
http://libraries.ou.edu/eresources/resolver.aspx?atitle=How+meteorologists+learn+to+forecast+the+weather%3A+Social+dimensions+of+complex+learning&author=LaDue%2C+Daphne+S.&volume=&issue=&spage=&date=2011&title=How+meteorologists+learn+to+forecast+the+weather%3A+Social+dimensions+of+complex+learning&issn= Dissertations & Theses @ University of Oklahoma; ProQuest Dissertations & Theses Full Text database.

LeClerc, J., & Joslyn, S. (2015). The Cry Wolf Effect and Weather□Related Decision Making. *Risk analysis, 35*(3), 385-395.

Lehmann, A., Schumann, H., Staadt, O., & Tominski, C. (2011). Physical navigation to support graph exploration on a large high-resolution display *Advances in Visual Computing* (pp. 496-507): Springer.

Levin, S., Sauer, L., Kelen, G., Kirsch, T., Pham, J., Desai, S., & France, D. (2012). Situation awareness in emergency medicine. *IIE Transactions on Healthcare Systems Engineering, 2*(2), 172-180. doi:10.1080/19488300.2012.684739

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making, 14*(5), 331-352. doi:10.1002/bdm.381

Lipshitz, R., & Strauss, O. (1997). Coping with Uncertainty: A Naturalistic Decision-Making Analysis. *Organizational Behavior and Human Decision Processes, 69*(2), 149-163. doi:http://dx.doi.org/10.1006/obhd.1997.2679

Loft, S., Morrell, D. B., & Huf, S. (2013). Using the situation present assessment method to measure situation awareness in simulated submarine track management. *International Journal of Human Factors and Ergonomics, 2*(1), 33-48.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated Choice Methods*: Cambridge University Press.

Lowe, R. K. (2004). Interrogation of a dynamic visualization during learning. *Learning and Instruction, 14*(3), 257-274.

Lowe, R. K. (2008). Learning from Animation: Where to Look, When to Look. In R. K. Lowe & W. Schnotz (Eds.), *Learning with Animation: Research Implications for Design*.

Ma, R., & Kaber, D. B. (2005). Situation awareness and workload in driving while using adaptive cruise control and a cell phone. *International Journal of Industrial Ergonomics, 35*(10), 939-953.

Mays, L. W. (2010). *Water Resources Engineering*: John Wiley & Sons.

Malakis, S., & Kontogiannis, T. (2013). A sensemaking perspective on framing the mental picture of air traffic controllers. *Applied Ergonomics, 44*(2), 327-339. doi: http://dx.doi.org/10.1016/j.apergo.2012.09.003

McIlroy, R. C., & Stanton, N. A. (2011). Getting past first base: Going all the way with Cognitive Work Analysis. *Applied Ergonomics, 42*(2), 358-370. doi:http://dx.doi.org/10.1016/j.apergo.2010.08.006

McNicol, D. (2005). *A primer of signal detection theory*: Psychology Press.

Minotra, D., & Burns, C. M. (2015). Finding Common Ground: Situation Awareness and Cognitive Work Analysis. *Journal of Cognitive Engineering and Decision Making, 9*(1), 87-89. doi:10.1177/1555343414555159

Montello, D. R., & Freundschuh, S. (2005). Cognition of geographic information. *A research agenda for geographic information science*, 61-91.

Moore, K. (2009). *Comparison of Eye Movement Data to Direct Measures of Situation Awareness for Development of a Novel Measurement Technique in Dynamic, Uncontrolled Test Environments.* (Doctor of Philosophy (PhD)), Clemson University, All Dissertations. Retrieved from http://tigerprints.clemson.edu/all_dissertations/477 (477)

Moore, K., & Gugerty, L. (2010). Development of a Novel Measure of Situation Awareness: The Case for Eye Movement Analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(19), 1650-1654. doi:10.1177/154193121005401961

Morss, R. E., Demuth, J. L., Bostrom, A., Lazo, J. K., & Lazrus, H. (2015). Flash Flood Risks and Warning Decisions: A Mental Models Study of Forecasters, Public Officials, and Media Broadcasters in Boulder, Colorado. *Risk analysis, 35*(11), 2009-2028. doi:10.1111/risa.12403

Morss, R. E., & Ralph, F. M. (2007). Use of information by National Weather Service forecasters and emergency managers during CALJET and PACJET-2001. *Weather and Forecasting, 22*(3), 539-555.

Muniz, E., Stout, R., Bowers, C., & Salas, E. (1998). A methodology for measuring team situational awareness: situational awareness linked indicators adapted to novel tasks (SALIENT). *NATO human factors and medicine panel on collaborative crew performance in complex systems, Edinburgh, North Atlantic Treaties Organisation, Neuilly-sur-Seine*, 20-24.

Murphy, A. H. (1978). Hedging and the Mode of Expression of Weather Forecasts. *Bulletin of the American Meteorological Society, 59*(4), 371-373. doi:10.1175/1520-0477(1978)059<0371:HATMOE>2.0.CO;2

Murphy, A. H. (1993). What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting, 8*(2), 281-293. doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

Murphy, A. H., & Daan, H. (1984). Impacts of Feedback and Experience on the Quality of Subjective Probability Forecasts. Comparison of Results from the First and Second Years of the Zierikzee Experiment. *Monthly Weather Review, 112*(3), 413-423. doi:10.1175/1520-0493(1984)112<0413:IOFAEO>2.0.CO;2

Nadav-Greenberg, L., Joslyn, S. L., & Taing, M. U. (2008). The effect of uncertainty visualizations on decision making in weather forecasting. *Journal of Cognitive Engineering and Decision Making, 2*(1), 24-47.

National Climatic Data Center. (2014). *Storm Events Database*. Retrieved from: https://www.ncdc.noaa.gov/stormevents/

National Weather Service. (2011). *Weather Forecast Office Hydrologic Products Specification*. (National Weather Service Directive 10-922). Washington, D.C. Retrieved from http://www.nws.noaa.gov/directives/sym/pd01009022curr.pdf.

National Weather Service. (2014a). *Flash Flood Warning #14.*

National Weather Service. (2014b). *The Record Front Range and Eastern Colorado Floods of September 11-17, 2013.* Retrieved from http://www.nws.noaa.gov/om/assessments/pdfs/14colorado_floods.pdf.

National Weather Service. (2014c). *Service Assessment: May 2013 Oklahoma Tornadoes and Flash Floods*. Retrieved from http://www.nws.noaa.gov/om/assessments/pdfs/13oklahoma_tornadoes.pdf.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*: WH Freeman/Times Books/Henry Holt & Co.

Novak, D. R., Bright, D. R., & Brennan, M. J. (2008). Operational forecaster uncertainty needs and future roles. *Weather and Forecasting, 23*(6), 1069-1084.

Ooms, K., De Maeyer, P., & Fack, V. (2013). Study of the attentive behavior of novice and expert map users using eye tracking. *Cartography and Geographic Information Science, 41*(1), 37-54. doi:10.1080/15230406.2013.860255

Ooms, K., De Maeyer, P., Fack, V., Van Assche, E., & Witlox, F. (2012). Interpreting maps through the eyes of expert and novice users. *International Journal of Geographical Information Science, 26*(10), 1773-1788. doi:10.1080/13658816.2011.642801

Pagano, T. C., Wood, A. W., Ramos, M.-H., Cloke, H. L., Pappenberger, F., Clark, M. P., . . . Verkade, J. S. (2014). Challenges of Operational River Forecasting. *Journal of Hydrometeorology, 15*(4), 1692-1707. doi:10.1175/JHM-D-13-0188.1

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(2), 230-253. doi:10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140-160. doi:10.1518/155534308x284417

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still Vital After All These Years of Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 511-520. doi:10.1518/001872008x312198

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods, 162*(1), 8-13.

Pielke, R. A., Downton, M. W., & Barnard Miller, J. Z. (2002). *Flood damage in the United States, 1926-2000: a reanalysis of National Weather Service estimates*: University Corporation for Atmospheric Research Boulder, CO.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological review, 106*(4), 643-675. doi:10.1037/0033-295X.106.4.643

Pirolli, P., & Card, S. (2005). *The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis.* Paper presented at the Proceedings of international conference on intelligence analysis.

Pliske, R. M., Crandall, B., & Klein, G. (2004). Competence in Weather Forecasting. In K. Smith, J. Shanteau, & P. Johnson (Eds.), *Psychological Investigations of Competence in Decision Making*. Cambridge, UK: Cambridge University Press.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of human computer interaction, 1*, 211-219.

Quoetone, E. M., Andra, D. L., Bunting, W. F., & Jones, D. G. (2001). *Impacts of Technology and Situation Awareness on Decision Making: Operational Observations from National Weather Service Warning Forecasters During the Historic May 3 1999 Tornado Outbreak.* Paper presented at the Human Factors and Ergonomics Society 45th Annual Meeting, Minneapolis, MN.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata, Second Edition*: Stata Press.

Randel, J. M., Pugh, H. L., & Reed, S. K. (1996). Differences in expert and novice situation awareness in naturalistic decision making. *International Journal of Human-Computer Studies, 45*(5), 579-597. doi: http://dx.doi.org/10.1006/ijhc.1996.0068

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied, 14*(1), 36. doi:10.1037/1076-898X.14.1.36

Read, G. J. M., Salmon, P. M., Lenné, M. G., & Stanton, N. A. (2014). Designing sociotechnical systems with cognitive work analysis: putting theory back into practice. *Ergonomics, 58*(5), 822-851. doi: 10.1080/00140139.2014.980335

Rector, K. (2014, June 10). Flash flood watch in effect after heavy rains inundate College Park. *The Baltimore Sun*. Retrieved from http://www.baltimoresun.com/news/weather/weather-blog/bs-md-flooding-20140610-story.html

Ripberger, J. T., Silva, C. L., Jenkins-Smith, H. C., & James, M. (2014). The Influence of Consequence-Based Messages on Public Responses to Tornado Warnings. *Bulletin of the American Meteorological Society, 96*(4), 577-590. doi:10.1175/BAMS-D-13-00213.1

Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics, 52*(5), 512-523. doi:10.1080/00140130802379129

Salas, E., Fiore, S. M., & Letsky, M. P. (2013). *Theories of Team Cognition: Cross-Disciplinary Perspectives*: Taylor & Francis.

Salas, E., Prince, C., Baker, D. P., & Shrestha, L. (1995). Situation awareness in team performance: Implications for measurement and training. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 123-136.

Saldana, J. (2015). *The Coding Manual for Qualitative Researchers*: SAGE Publications.

Salmon, P. M., Stanton, N. A., Walker, G. H., Baber, C., Jenkins, D. P., McMaster, R., & Young, M. S. (2008). What really is going on? Review of situation awareness models for individuals and teams. *Theoretical Issues in Ergonomics Science, 9*(4), 297-323. doi:10.1080/14639220701561775

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics, 39*(3), 490-500. doi: http://dx.doi.org/10.1016/j.ergon.2008.10.010

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D. P., & Rafferty, L. (2010). Is it really better to share? Distributed situation awareness and its implications for collaborative system design. *Theoretical Issues in Ergonomics Science, 11*(1-2), 58-83. doi:10.1080/14639220903009953

Salmon, P. M., Stanton, N. A., & Young, K. L. (2012). Situation awareness on the road: review, theoretical and methodological issues, and future directions. *Theoretical Issues in Ergonomics Science, 13*(4), 472-492. doi: 10.1080/1463922X.2010.539289

Sarter, N. B., & Woods, D. D. (1991). Situation Awareness: A Critical But Ill-Defined Phenomenon. *The International Journal of Aviation Psychology, 1*(1), 45-57. doi:10.1207/s15327108ijap0101_4

Schreier, M. (2012). *Qualitative content analysis in practice*: Sage Publications.

Selcon, S. J., Taylor, R. M., & Koritsas, E. (1991). Workload or Situational Awareness?: TLX vs. SART for Aerospace Systems Design Evaluation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 35*(2), 62-66. doi: 10.1518/107118191786755706

Shah, P., & Freedman, E. G. (2011). Bar and Line Graph Comprehension: An interaction of top□down and bottom□up processes. *Topics in cognitive science, 3*(3), 560-578.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems, 33*(2), 111-126. doi:http://dx.doi.org/10.1016/S0167-9236(01)00139-7

Shneiderman, B. (1996). *The eyes have it: A task by data type taxonomy for information visualizations.* Paper presented at the Visual Languages, 1996. Proceedings., IEEE Symposium on.

Shrestha, A., Zhu, Y., & Miller, B. (2014). *Visualizing Uncertainty in Spatio-temporal data.* Paper presented at the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA).

Shu, Y., & Furuta, K. (2005). An inference method of team situation awareness based on mutual awareness. *Cognition, Technology & Work, 7*(4), 272-287. doi:10.1007/s10111-005-0012-x

Smallman, H. S., & Hegarty, M. (2007). *Expertise, spatial ability and intuition in the use of complex visual displays.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Smith, K., & Hancock, P. A. (1995). Situation Awareness Is Adaptive, Externally Directed Consciousness. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 137-148. doi:10.1518/001872095779049444

St John, M., Callan, J., Proctor, S., & Holste, S. (2000). *Tactical decision-making under uncertainty: Experiments I and II (No. TR-1821)*. Retrieved from Pacific Sciences and Engineering Group Inc. San Diego, CA: http://www.dtic.mil/dtic/tr/fulltext/u2/a378170.pdf

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers, 31*(1), 137-149.

Stanton, N. A., Salmon, P. M., & Walker, G. H. (2015). Let the Reader Decide: A Paradigm Shift for Situation Awareness in Sociotechnical Systems. *Journal of Cognitive Engineering and Decision Making, 9*(1), 44-50. doi:10.1177/1555343414552297

Stanton, N. A., Salmon, P. M., Walker, G. H., & Jenkins, D. (2008). Genotype and phenotype schemata and their role in distributed situation awareness in collaborative systems. *Theoretical Issues in Ergonomics Science, 10*(1), 43-68. doi:10.1080/14639220802045199

Stanton, N. A., Salmon, P. M., Walker, G. H., & Jenkins, D. P. (2009). Is situation awareness all in the mind? *Theoretical Issues in Ergonomics Science, 11*(1-2), 29-40. doi:10.1080/14639220903009938

Stanton, N. A., Stewart, R., Harris, D., Houghton, R. J., Baber, C., McMaster, R., . . . Green, D. (2006). Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology. *Ergonomics, 49*(12-13), 1288-1311. doi:10.1080/00140130600612762

Stewart, T. R., Heideman, K. F., Moninger, W. R., & Reagan-Cirincione, P. (1992). Effects of improved information on the components of skill in weather forecasting. *Organizational Behavior and Human Decision Processes, 53*(2), 107-134. doi:http://dx.doi.org/10.1016/0749-5978(92)90058-F

Stuart, N. A., Market, P. S., Telfeyan, B., Lackmann, G. M., Carey, K., Brooks, H. E., . . . Reeves, K. (2006). The Future of Humans in an Increasingly Automated Forecast Process. *Bulletin of the American Meteorological Society, 87*(11), 1497-1502. doi:10.1175/BAMS-87-11-1497

Stuart, N. A., Schultz, D. M., & Klein, G. (2007). Maintaining the Role of Humans in the Forecast Process: Analyzing the Psyche of Expert Forecasters. *Bulletin of the American Meteorological Society, 88*(12), 1893-1898. doi:10.1175/BAMS-88-12-1893

Sturre, L., Chiappe, D., Vu, K.-P. L., & Strybel, T. Z. (2015). Using Eye Movements to Test Assumptions of the Situation Present Assessment Method. In S. Yamamoto

(Ed.), *Human Interface and the Management of Information. Information and Knowledge in Context: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II* (pp. 45-52). Cham: Springer International Publishing.

Taylor, R. M. (1990). Situational Awareness Rating Technique(SART): The development of a tool for aircrew systems design. *AGARD, Situational Awareness in Aerospace Operations 17 p(SEE N 90-28972 23-53)*.

Trafton, J. G. (2004). Dynamic Mental Models in Weather Forecasting. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48*(3), 311-314. doi:10.1177/154193120404800308

Trafton, J. G., & Hoffman, R. (2007). *Computer-aided visualization in meteorology*: Lawrence Erlbaum.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: extracting and generating information from complex visualizations. *International Journal of Human-Computer Studies, 53*(5), 827-850. doi:http://dx.doi.org/10.1006/ijhc.2000.0419

Trickett, S. B., & Trafton, J. G. (2006). Toward a Comprehensive Model of Graph Comprehension: Making the Case for Spatial Cognition. In D. Barker-Plummer, R. Cox, & N. Swoboda (Eds.), *Diagrammatic Representation and Inference* (Vol. 4045, pp. 286-300): Springer Berlin Heidelberg.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131. doi: 10.1126/science.185.4157.1124

van de Merwe, K., Oprins, E., Eriksson, F., & van der Plaat, A. (2012). The Influence of Automation Support on Performance, Workload, and Situation Awareness of Air Traffic Controllers. *The International Journal of Aviation Psychology, 22*(2), 120-143. doi:10.1080/10508414.2012.663241

van de Merwe, K., van Dijk, H., & Zon, R. (2012). Eye Movements as an Indicator of Situation Awareness in a Flight Simulator Experiment. *The International Journal of Aviation Psychology, 22*(1), 78-95. doi:10.1080/10508414.2012.635129

van Winsen, R., & Dekker, S. W. A. (2015). SA Anno 1995: A Commitment to the 17th Century. *Journal of Cognitive Engineering and Decision Making, 9*(1), 51-54. doi:10.1177/1555343414557035

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*: CRC Press.

Walker, G. H., Gibson, H., Stanton, N. A., Baber, C., Salmon, P., & Green, D. (2006). Event analysis of systemic teamwork (EAST): a novel integration of ergonomics methods to analyse C4i activity. *Ergonomics, 49*(12-13), 1345-1369. doi: 10.1080/00140130600612846

Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3): Sage.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization science, 16*(4), 409-421.

Wickens, C. D. (2008). Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 397-403. doi:10.1518/001872008x288420

Wickens, C. D. (2015). Situation Awareness: Its Applications Value and Its Fuzzy Dichotomies. *Journal of Cognitive Engineering and Decision Making, 9*(1), 90-94. doi:10.1177/1555343414564571

Wickens, C. D., & Carswell, C. M. (1997). Information processing. *Handbook of human factors and ergonomics, 2*, 89-122.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., & Zheng, S. (2008). Attention-situation awareness (A-SA) model of pilot error. *Human performance modeling in aviation*, 213-239.

Wilson, D., & Sperber, D. (2002). Relevance theory. *Handbook of pragmatics*.

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., . . . Whitehead, P. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research, 47*(5).

Yu, C. S., Wang, E. M., Li, W. C., & Braithwaite, G. (2014). Pilots' visual scan patterns and situation awareness in flight operations. *Aviat Space Environ Med, 85*(7), 708-714.

Appendix A: Logistic Regression Odds Ratios from Chapter 3

	Marginal Effects	*Conditional Effects*
	Ordinary Logistic	**Random int. logistic**
	OR (95% CI)	**OR (95% CI)**
Fixed part		
Intercept	6.50 (4.60, 9.50)	6.50 (4.60, 9.50)
exp(β_2) [size]	0.011 (0.003, 0.032)	0.011 (0.003, 0.032)
exp(β_3) [threat]	0.15 (0.089, 0.23)	0.15 (0.089, 0.23)
exp(β_4) [display]	0.29 (0.18, 0.47)	0.29 (0.18, 0.47)
exp(β_5) [size*threat]	206.87 (63.16, 946.22)	206.87 (63.16, 946.22)
exp(β_6) [size*display]	17.39 (5.21, 80.19)	17.39 (5.21, 80.19)
exp(β_7) [threat*display]	11.92 (5.93, 24.37)	11.92 (5.93, 24.37)
exp(β_8) [size*threat*display]	0.043 (0.008, 0.18)	0.043 (0.008 0.18)
Random Part		
ψ	--	0
ρ	--	0
Goodness of Fit		
Log likelihood	-630.80	-630.80
Hosmer-Lemeshow	1.00	

Appendix B: Guidance Products Available in HWT-Hydro 2014

Experimental Products

Experimental Models	CREST Maximum Return Period
	HRRR-Forced CREST
	CREST Soil Moisture
	CREST Streamflow
	SAC-SMA Soil Moisture
	SAC-SMA Streamflow
Precipitable Water (PW)	Precipitable Water Analysis (RAOBs)
	Precipitable Water Percentile (RAOBs)
	Precipitable Water Analysis (RAP)
	Precipitable Water Percentile (RAP)
Quantitative Precipitation Estimate (QPE)	MRMS QPE
Quantitative Precipitation Forecast (QPF)	MRMS QPF
Flash Flood Guidance Ratio (FFG)	QPE to Flash Flood Guidance Ratio
	QPF to Flash Flood Guidance Ratio
Average Recurrence Interval (ARI)	Precipitation Return Period (QPE)
	Precipitation Return Period (QPF)

Operational Products

FFA/FFW/LSRs	Flash Flood Advisories
	Flash Flood Watch (FFA)
	Flash Flood Warnings (FFW)
	Experimental Flash Flood Warnings
	Flood Warnings
	Local Storm Reports (LSRs)
Radar	MRMS Seamless Hybrid-Scan Reflectivity
	MRMS Quality-Controlled Composite Reflectivity
Satellite	Infrared (IR) Window
	Water Vapor Satellite
	Visible Satellite
	3.9u, 13u, 11u-3.9u, 11u-13u, 3.7u, 3.7u-13u
	WV/IR

Ensembles and Data Assimilation	ECMWF-HiRes Model
	GFS20
	GFS
	HiResW-ARW-East and West
	HiResW-NMM-East and West
	High Resolution Rapid Refresh Model (HRRR)
	LAPS
	NAM12, NAM40, NAM80
	RAP13, RAP40
	SREF
	UKMET Ensemble
Observations	Surface Plots
	METAR Station Plots
	Synoptic Plots
Overlays	State Boundaries and Names
	County Boundaries and Names
	Rivers and Streams
	Lakes
	River Drainage Basins
	Cities
	County Warning Area Boundaries
	River Forecast Center Boundaries
	Interstates and US Highways
	Railroads
	High Resolution Topographic Imagery

Appendix C: Portion of a Screen Recording Transcript from HWT-Hydro

Product	*Start Time*	*End Time*
Web Maximum Return Period CREST	2:28:28	2:32:28
FLASH Surface MRMS Seamless Hybrid-Scan Reflectivity	2:32:28	2:32:51
FLASH Surface 1-hr Precipitation Return Period (Forecast)	2:32:28	2:32:51
FLASH Surface 3-hr Precipitation Return Period (Forecast)	2:32:28	2:32:51
FLASH Surface 6-hr Precipitation Return Period (Forecast)	2:32:28	2:32:51
FLASH Surface HRRR-Forced CREST	2:32:51	2:32:56
Local Storm Reports	2:32:56	2:33:06
Flood Advisories	2:32:56	2:33:06
Flood Warnings	2:32:56	2:33:06
Flash Flood Warnings	2:32:56	2:33:06
Experimental Flash Flood Warnings	2:32:56	2:33:06
FLASH Surface MRMS Seamless Hybrid-Scan Reflectivity	2:32:56	2:33:44
FLASH Surface 1-hr Precipitation Return Period (Forecast)	2:33:06	2:33:44
FLASH Surface 3-hr Precipitation Return Period (Forecast)	2:33:06	2:33:44
FLASH Surface 6-hr Precipitation Return Period (Forecast)	2:33:06	2:33:44
Flood Advisories	2:33:44	2:34:31
Flood Warnings	2:33:44	2:34:31
Flash Flood Warnings	2:33:44	2:34:31
Experimental Flash Flood Warnings	2:33:44	2:34:31
FLASH Surface MRMS Seamless Hybrid-Scan Reflectivity	2:33:44	2:35:33
FLASH Surface 1-hr MRMS Radar-Only QPE to FFG Ratio	2:33:44	2:34:09
FLASH Surface 3-hr MRMS Radar-Only QPE to FFG Ratio	2:33:44	2:34:09
FLASH Surface 6-hr MRMS Radar-Only QPE to FFG Ratio	2:33:44	2:34:09
Local Storm Reports	2:34:09	2:34:31
FLASH Surface 1-hr Precipitation Return Period (Forecast)	2:34:31	2:34:36
FLASH Surface 3-hr Precipitation Return Period (Forecast)	2:34:31	2:34:36

Appendix D: Focus Group Questions in HWT-Hydro 2014

General Forecasting Experience

1. Please describe your approach in issuing forecasts and warnings this week. What tools and products do you usually use to guide your decisions when issuing flash flood watches and warnings? (If issuing watches and warnings is not part of your current job, what products have you used in the past or during training exercises?)
2. What challenges or difficulties did you encounter while forecasting this week?

Uncertainty, Probability, and Confidence

3. What role does your personal confidence (in terms of what the models predict, what your background experience indicates, your extrapolation of future events, etc.) play in producing the flash flood watches? Flash flood warnings?

We are interested in knowing more about your opinions about categorization of flash floods:

4. How did issuing attributes of severity for watches and (nuisance v. major) enable you to communicate threat information?
5. What did you find helpful about the categorization? What would you change about the categorization?
6. Did participating in HWT-Hydro 2014 affect how you view probabilities in flash flood forecasting? How?
7. When assigning uncertainty estimates to the magnitudes, what factors affected your decisions? For example, were there any cases where you were more or less likely to issue a warning based on factors like geography?

We would now like to ask some questions about watches and lead times.

8. What are your thoughts on the current paradigm for issuing flash flood watches?
9. What benefits and challenges do producing flash flood watches with long lead times afford? Can you think of any lead time that would to absolutely too long?
10. What benefits and challenges do producing flash flood watches with short lead times afford?
11. What kind of products would be most useful to you at the flash flood watch scale? Warning scale?

12. Do you have any further comments about your experiences with issuing watches and warnings during the past week?

Appendix E: List of Probes Used in Chapter 5

SA Level	Probe	Administration (in Qualtrics)	Scoring Method
1	Using the provided (blank) map, point out the area(s) that received the highest [return period / streamflow / QPE-to-FFG ratio] values in the past two hours.	Heat Map	+1 point for selecting an area within the correct area +0.5 point for selecting an area within the correct county +0.25 points for selecting an area within an adjacent county +0 point for being outside of range or a miss
	In [units] over the past two hours, what was the peak [metric, e.g. return period] value reached?	Manual entry	+1 for being within ± 20 units from target) +0 for being outside range
2	Using the provided (blank) map, point out all the areas that have conditions associated with flash flooding at the most recent timestamp.	Heat map	+1 point for selecting an area within the correct area +0.5 point for selecting an area within the correct county +0.25 points for selecting an area within an adjacent county +0 point for being outside of range or a miss
	Using the provided map, click which polygons highlight areas with conditions associated with flash flooding at the current time.	Hot spot	+1 for each correct identification +0 for not warning on either areas
3	Using the provided (blank) map, point out all the areas you would expect to be under greatest risk for flash flooding in the next two hours.	Heat map	+1 point for selecting an area within the correct area +0.5 point for selecting an area

			within the correct county +0.25 points for selecting an area within an adjacent county +0 point for being outside of range or a miss
	In hours, indicate the lengths of valid time you would assign to the warnings in this region.	Manual entry	+1 point for matching time to stream gage falling below flood stage +0.5 point for ± 2 hours past stream gage falling below flood stage +0 for being outside range

Appendix F: Example Briefing Used in Chapter 5

```
URGENT - IMMEDIATE BROADCAST REQUESTED
FLOOD WATCH
NATIONAL WEATHER SERVICE NORTHERN INDIANA
826 PM EDT MON JUL 13 2015

INZ003-012-013-015-020-022>027-032>034-140800-
/O.EXT.KIWX.FF.A.0005.000000T0000Z-150714T0800Z/
/00000.0.ER.000000T0000Z.000000T0000Z.000000T0000Z.OO/
LA PORTE-STARKE-PULASKI-FULTON IN-WHITE-CASS IN-MIAMI-
WABASH-HUNTINGTON-WELLS-ADAMS-GRANT-BLACKFORD-JAY-
INCLUDING THE CITIES OF...MICHIGAN CITY...LA PORTE...
KNOX...NORTH JUDSON...BASS
LAKE...WINAMAC...FRANCESVILLE...
MEDARYVILLE...ROCHESTER...AKRON...MONTICELLO...BROOKSTON..
.MONON...LOGANSPORT...ROYAL CENTER...PERU...GRISSOM
AFB...MEXICO...WABASH...NORTH MANCHESTER...HUNTINGTON...
ROANOKE...BLUFFTON...OSSIAN...DECATUR...BERNE...MARION...
GAS CITY...UPLAND...HARTFORD CITY...MONTPELIER...
PORTLAND...DUNKIRK
826 PM EDT MON JUL 13 2015

...FLASH FLOOD WATCH NOW IN EFFECT UNTIL 4 AM EDT /3 AM
CDT/TUESDAY...

THE FLASH FLOOD WATCH IS NOW IN EFFECT FOR

* A PORTION OF NORTHERN INDIANA...INCLUDING THE FOLLOWING
  AREAS...ADAMS...BLACKFORD...CASS IN...FULTON IN...
GRANT... HUNTINGTON...JAY...LA PORTE...MIAMI...
PULASKI...STARKE... WABASH...WELLS AND WHITE.

* UNTIL 4 AM EDT /3 AM CDT/ TUESDAY

* FIRST ROUND OF STORMS THIS MORNING DROPPED BETWEEN AN
INCH AND AN INCH AND A HALF ACROSS MOST OF THE FORECAST
AREA.

* FLOODING WILL QUICKLY OCCUR ANYWHERE STORMS DEVELOP THIS
EVENING INTO THE EARLY PORTION OF THE OVERNIGHT HOURS WITH
ANOTHER 1 TO 2 INCHES EXPECTED FROM ANY OF THESE STORMS. A
FEW ISOLATED LOCATIONS MAY SEE 3 OR MORE INCHES OF RAIN
TONIGHT...ESPECIALLY IF TRAINING STORMS DEVELOP.
```

```
URGENT - IMMEDIATE BROADCAST REQUESTED
FLOOD WATCH
NATIONAL WEATHER SERVICE INDIANAPOLIS IN
```

347 PM EDT MON JUL 13 2015

...FLASH FLOODING POSSIBLE THROUGH TONIGHT...

.ADDITIONAL ROUNDS OF STRONG TO SEVERE THUNDERSTORMS ARE
EXPECTED TO DEVELOP THROUGH LATE TONIGHT WITH THE
POTENTIAL TO PRODUCE SIGNIFICANT RAINFALL AMOUNTS.
WIDESPREAD RAINFALL OF 2 TO 5 INCHES HAS OCCURRED IN THE
LAST SIX DAYS... WITH POCKETS AS HIGH AS NEARLY 9 INCHES
OF RAINFALL. ADDITIONAL SIGNIFICANT RAINFALL WILL BE
LIKELY TO CAUSE FLASH FLOODING IN THESE AREAS.

INZ021-028>031-035>049-051>057-060>065-067>072-140400-
/O.CAN.KIND.FF.A.0005.000000T0000Z-150714T0900Z/
/O.NEW.KIND.FF.A.0006.150713T2000Z-150714T1200Z/
/00000.0.ER.000000T0000Z.000000T0000Z.000000T0000Z.OO/
CARROLL-WARREN-TIPPECANOE-CLINTON-HOWARD-FOUNTAIN-
MONTGOMERY-BOONE-TIPTON-HAMILTON-MADISON-DELAWARE-
RANDOLPH-VERMILLION-PARKE-PUTNAM-HENDRICKS-MARION-HANCOCK-
HENRY-VIGO-CLAY-OWEN-MORGAN-JOHNSON-SHELBY-RUSH-SULLIVAN-
GREENE-MONROE-BROWN-BARTHOLOMEW-DECATUR-KNOX-DAVIESS-
MARTIN-LAWRENCE-JACKSON-JENNINGS-INCLUDING THE CITIES
OF...LAFAYETTE...FRANKFORT...KOKOMO...CRAWFORDSVILLE...
ANDERSON...MUNCIE...INDIANAPOLIS...TERRE HAUTE...
SHELBYVILLE...BLOOMINGTON...COLUMBUS...VINCENNES...
BEDFORD...SEYMOUR

347 PM EDT MON JUL 13 2015

...FLASH FLOOD WATCH IN EFFECT THROUGH TUESDAY MORNING...

THE NATIONAL WEATHER SERVICE IN INDIANAPOLIS HAS ISSUED A

* FLASH FLOOD WATCH FOR A PORTION OF INDIANA...INCLUDING
THE FOLLOWING AREAS...BARTHOLOMEW...BOONE...BROWN...
CARROLL...CLAY...CLINTON...DAVIESS...DECATUR...DELAWARE...
FOUNTAIN...GREENE...HAMILTON...HANCOCK...HENDRICKS...
HENRY...HOWARD...JACKSON...JENNINGS...JOHNSON...KNOX...
LAWRENCE...MADISON...MARION...MARTIN...MONROE...
MONTGOMERY...MORGAN...OWEN...PARKE...PUTNAM...RANDOLPH...
RUSH...SHELBY...SULLIVAN...TIPPECANOE...TIPTON...VERMILLIO
N...VIGO AND WARREN.

* THROUGH TUESDAY MORNING
* WIDESPREAD SIGNIFICANT RAINFALL HAS OCCURRED OVER THE
LAST SIX DAYS...WITH ADDITIONAL SIGNIFICANT RAINFALL
POSSIBLE TONIGHT. AREAS RECEIVING SIGNIFICANT RAINFALL
WILL BE LIKELY TO EXPERIENCE FLASH FLOODING.