Investigating the Relationship between Eye Movements and Situation Awareness in Weather Forecasting

Elizabeth M. Argyle[1], Jonathan J. Gourley[2], Ziho Kang[3], and Randa L. Shehab[3]

[1] Human Factors Research Group, University of Nottingham, Nottingham, NG7 2RD, UK

[2] NOAA, National Severe Storms Laboratory, Norman, OK, 73072, USA

[3] School of Industrial and Systems Engineering, University of Oklahoma, Norman, OK, 73079, USA

29 January 2020

Accepted to Applied Ergonomics

This manuscript is the author's accepted manuscript. The published full-text can be found at

https://doi.org/10.1016/j.apergo.2020.103071

ABSTRACT

Physiological indicators, including eye tracking measures, may provide insight into human decision making and cognition in many domains, including weather forecasting. Situation awareness (SA), a critical component of forecast decision making, is commonly conceptualized as the degree to which information is perceived, understood, and projected into a future context. Drawing upon recent applications of eye tracking in the study of forecaster decision making, we investigate the relationship among eye movement measures, automation, and SA assessed through a freeze probe assessment method. In addition, we explore the relationship between an automated forecasting decision aid use and information seeking behavior.

In this study, a sample of professional weather forecasters completed a series of tasks, informed by a set of forecasting decision aids, and with variable access to an experimental automated tool, while an eye tracking system captured data related to eye movements and information usage. At the end of each forecasting task, participants responded to a set of questions related to the environmental situation in the framework of a survey-based assessment technique in order to assess their level of situation awareness. Regression analysis revealed a moderate relationship between the SA measure and eye tracking metrics, supporting the hypothesis that eye tracking may have utility in assessing SA. The results support the use of eye tracking in the assessment of specific and measurable attributes of the decision-making process in weather forecasting. The findings are discussed in light of potential benefits that eye tracking could bring to human performance assessment as well as decision-making research in the forecasting domain.

Keywords: situation awareness, human factors, assessment, eye tracking, weather forecasting

1. INTRODUCTION

The weather forecasting domain exemplifies the definition of a sociotechnical system with competing cognitive demands, and maintaining situation awareness (SA) is considered to be one critical aspect of forecasting (Quoetone et al. 2001; Endsley & Hoffman, 2002). The United States' National Weather Service's definition aligns with the Endsley 1995 model which frames SA in three stages: perception, comprehension, and projection of the current environmental state into a future state (Endsley, 1995b). Forecasters perceive and integrate information into a situational model upon which they base decisions, and should the model be inaccurate or incomplete, the forecast's spatial accuracy and timeliness may suffer, increasing the potential for negative societal impacts (Quoetone et al., 2001). Furthermore, as workplaces become increasingly automated, it is growing ever more important to ensure that automation is designed to support end users in developing and maintaining their situation awareness while reducing cognitive demand and out-of-the-loop decision making. Automation in the workplace affects task performance and decision-making within a variety of domains (Dao et al., 2009; Endsley & Kiris, 1995). In the weather forecasting domain, decision support automation may have potential to reduce time-consuming situation assessment activities within the forecasting process. While some evidence indicated that an early version of a forecasting automation tool did not significantly reduce forecaster workload (Karstens et al., 2015), this work hypothesized that automation would affect situation assessment and forecasters' levels of SA during a flash flood forecasting task.

In the search for objective, minimally distracting measures of human cognition, eye movement analysis has emerged as one such innovative solution (Moore & Gugerty, 2010; van de Merwe et al., 2012; Sturre et al., 2015). Traditionally, eye tracking has been employed in usability and human-computer interaction studies (Poole & Ball, 2006), but recent work has explored the technique's efficacy for cognitive-behavioral assessment (Wilson et al, 2016; Yu et al., 2016; McClung & Kang, 2016). Studying air traffic controllers, Moore and Gugerty (2010) identified an inverse relationship between error rate and the number of eye fixations, while also observing that eye fixations within an area of interest was the strongest predictor of overall awareness. Building upon previous research in this area, this work investigates the relationship between eye movements and situation awareness (SA) in weather forecasting. As physiological sensing technologies reach maturity, studies have demonstrated that such techniques may offer insight into aspects of human performance and cognition, such as workload (Heine et al., 2017; Marinescu et al., 2018) and SA (Moore & Gugerty, 2010; Catherwood et al., 2014; Ikuma et al., 2014; Wilson et al. 2016). Being able to assess SA in situ, objectively and with minimal disruption, could contribute to improved human-system integration, training development, and an enhanced understanding of behavioral patterns

in forecasting. However, before this can be implemented in an operational setting, it is necessary for research to verify and validate physiological methods against these purposes.

Physiological measures can complement a number of SA assessment techniques, ranging from probe-based techniques, subjective rating approaches, and observational methods. Probe-based methods, such as the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995a) and the Situation Present Assessment Method (SPAM; Pierce, 2012; Loft, Morrell, & Huf, 2013) evaluate SA against a predefined ground truth and reflect it as accuracy-based and time-based measures, respectively. However, while several probe- and rating-based approaches have been validated in laboratory settings, some work has found these to be effective only in controlled environments (Moore & Gugerty, 2010). Compared to probe-based or ratings-based techniques, eye tracking may provide a more direct way to evaluate SA (Adams et al., 1995). Furthermore, in environments where maintaining SA is critical, measures that reduce workplace interruption may improve assessment accuracy and operational safety.

Several studies have indicated that eye tracking may assess SA in other domains (Moore & Gugerty, 2010; van de Merwe et al., 2012; Yu et al., 2014), but to the authors' knowledge, the current work is the first to examine the relationship in the weather forecasting domain. Secondly, we explored differences in information usage when forecasters were exposed to an automated decision support tool. In previous work, Wilson et al. (2016) first demonstrated that forecaster eye movement data could reveal components of the forecast decision making process in terms of information scanning. Similarly, Wilson et al. (2017) employed eye tracking to capture scanning behavior during a workload assessment comparing multiple radar analysis tasks. Complementing and extending previous research, the current work investigated the relationship between eye tracking data and situation awareness in order to assess its utility for future in situ evaluations. Furthermore, the combination of probe-based assessment and eye tracking contributes to a deeper understanding of how graphical attention-directing mechanisms affect forecaster SA and information scanning.

In line with these hypothesis, the primary objective of this research was to explore the relationship between automation usage, eye movements, and SA (RQ1, RQ2):

RQ1: How is SA influenced by forecasting automation at different levels during a weather forecasting task?

RQ2: What is the effect of automation availability and decision aid use on total fixation duration, mean fixation count, and inter-AOI glances?

In addition, in order to assess SA during automation use, this work employed eye tracking 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 assessment. Existing work suggests a positive relationship between eye tracking measures, SA, and decision making, but this has been explored in weather forecasting to a much lesser degree (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 suitability of eye tracking measures in relation to predicting SA in flash flood forecasting (RQ3):

RQ3: To what degree are eye tracking measures able to predict situation awareness?

2. METHOD

2.1 Participants

Eighteen expert participants were recruited from NWS Forecast Offices, River Forecaster Centers, and other NWS Centers in the central United States. All participants had experience working in professional forecasting with a mean of 19.3 years of professional weather forecasting experience ($\sigma = 7.95$) and a mean of 14.6 years of hydrologic forecasting experience ($\sigma = 7.25$). Due to the nature of the study, it was preferable, though not mandatory, for participants to have specialized in hydrologic forecasting. The study was approved by the University of Oklahoma's Institutional Review Board.

2.2 Experimental Design

The study employed a two-factor, within-subjects design to assess the relationship among eye tracking, automation, and SA in flash flood forecasting. The primary independent variable was the availability of a forecasting decision support automation, which was presented in two levels (available or unavailable), while the secondary independent variable was the forecasting guidance product (Table 1). The dependent variables included an estimate of participant SA and eye tracking measures including number of fixations, total fixation duration and the proportion of inter-AOI glances captured in an area of interest (AOI) analysis. In this type of analysis, the researcher defines geometric regions of the display (AOIs), and software then quantifies eye movement metrics that occur within the areas (Poole and Ball, 2006). An additional benefit of AOI analysis is that it permits the researcher to compare participant interactions with different components of a display.

The eye tracking dependent variables (total fixation duration per AOI and mean number of fixations per AOI) were dependent upon assignment of Areas of Interest (AOIs), corresponding to the forecast guidance products, described in Table 1. Total fixation duration refers to the amount of time a participant fixated within an AOI over the entire recording period. The number of fixations variable measures the frequency of fixations within an AOI. This parameter can 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). Using the Tobii Pro Studio analysis software, four AOIs were evaluated, as shown in Figure 1.

Table 1. Forecasting guidance products available to participants

Decision Aid (Abbreviation)	Units	Description
CREST Unit Streamflow (USF)	$m^3s^{-1}km^{-2}$ (cubic meters per second per square kilometer)	Simulated surface water flows normalized by drainage area, selected from a span of 0.5 – 6 hours after the valid time
Average Recurrence Interval (ARI)	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 (decibel relative to Z)	Mosaic of reflectivity values measured by MRMS radars across the CONUS.

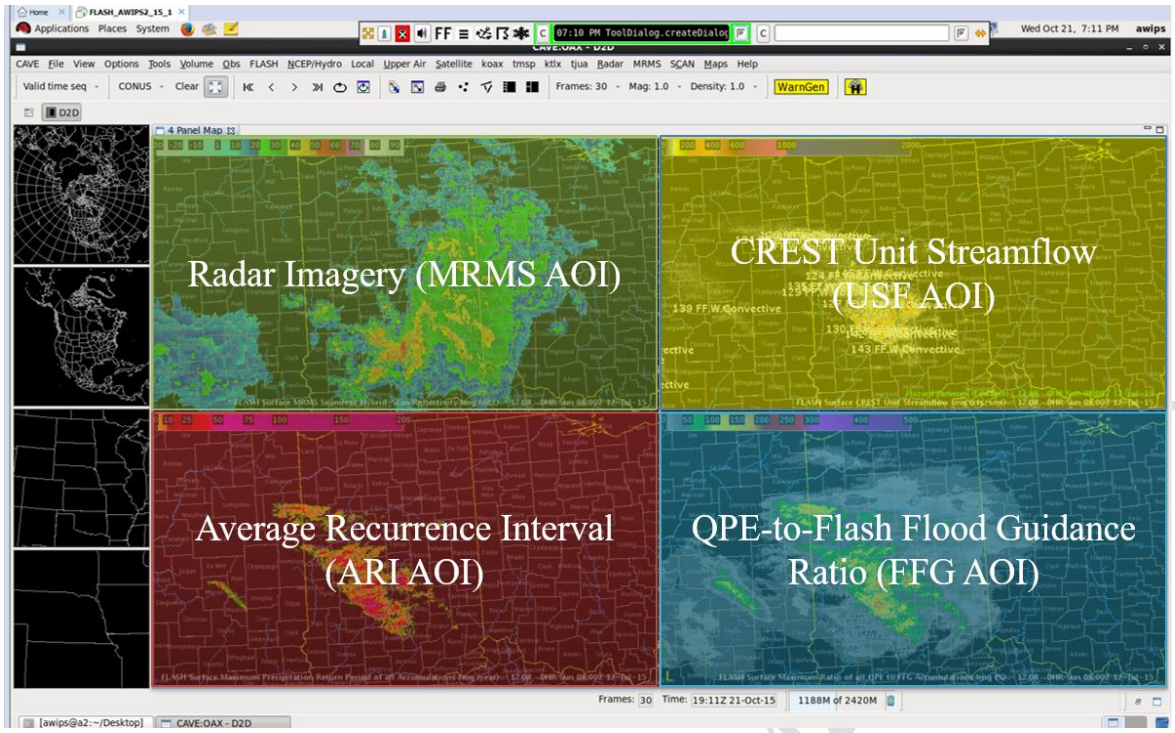

Figure 1. Areas of interest (AOIs) used during the analysis

2.3 Materials

2.3.1 Scenarios. Three scenarios were selected from flash flood reports in the NOAA Storm Data publication between May and July 2015 (available online from https://www.ncdc.noaa.gov/IPS/sd/sd.html). Each scenario consisted of a set of dynamic forecasting visualizations spanning two consecutive four-hour timeframes coinciding with the valid times of operational flash flood warnings issued by local National Weather Service (NWS) Forecast Offices. The timeframes involved real-world flooding events that took place in New Jersey (31 May 2015 21:30 UTC to 1 June 2015 01:30 UTC), southern Indiana (12 July 2015 06:00 to 10:00 UTC), and northern Kentucky (14 July 2015 19:00 to 23:00 UTC). In addition to operational warnings, historical reports from United States Geological Survey (USGS) stream gages were assessed during the timeframes. Selecting timeframes 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.

2.3.2 Forecast Guidance Products. The guidance products, described in Table 1, were presented with the Advanced Weather Interactive Processing System II (AWIPS-II), a computer visualization display platform used operationally within the NWS. The display was divided into four panels with one visualization per quadrant. Participants could not change the arrangement of the panels, the type of guidance products, or the visualizations' color palettes. However, they were permitted to zoom within and pan across the visualizations.

It was anticipated that participants would have varying levels of experience with the guidance products. For example, Multi-Radar/Multi-Sensor composite reflectivity radar imagery (MRMS) and the Quantitative Precipitation Estimate to Flash Flood Guidance (QPE-to-FFG) ratio product are both frequently used in operational environments, and as such should be familiar to forecasters. Conversely, the Average Recurrence Interval (ARI) and unit streamflow (USF) products were in-development and was not included in operational NWS information displays at the time of this work. Thus, it was anticipated that fewer forecasters would have experience using these products. For a technical discussion of the forecast guidance products, see Gourley et al. (2017).

2.3.3 Automation Design. Simulating an attention-directing guidance product, the present automation was a single-input threshold-exceedance parameter. Here, the QPE-to-FFG Ratio visualization provided the underlying model input, and the exceedance algorithm was based on a 1.0 ratio threshold. In operation, the automated product constructed white polygons around geographic regions of 10 square kilometers or more in which all model grid cells met or exceeded the 1.0 ratio threshold. As the automated algorithm was based on an initial prototype and not an operationally ready system, it is important to note that this study did not intend to test the effectiveness of the system.

The automated polygons were visualized within the upper-right quadrant of the display within the unit streamflow visualization. Participants were permitted to toggle between the unit streamflow map and the polygons to improve the visibility of the polygons, or could also overlay the two products; an example where the underlying unit streamflow visualization is hidden is shown in Figure 2. As placing the polygons over the QPE-to-FFG Ratio guidance product would have merely resulted in highlighting the regions already represented as “at-risk” by the color scale, this placement was expected to allow participants to assess the overlap in risk between the two decision aids.

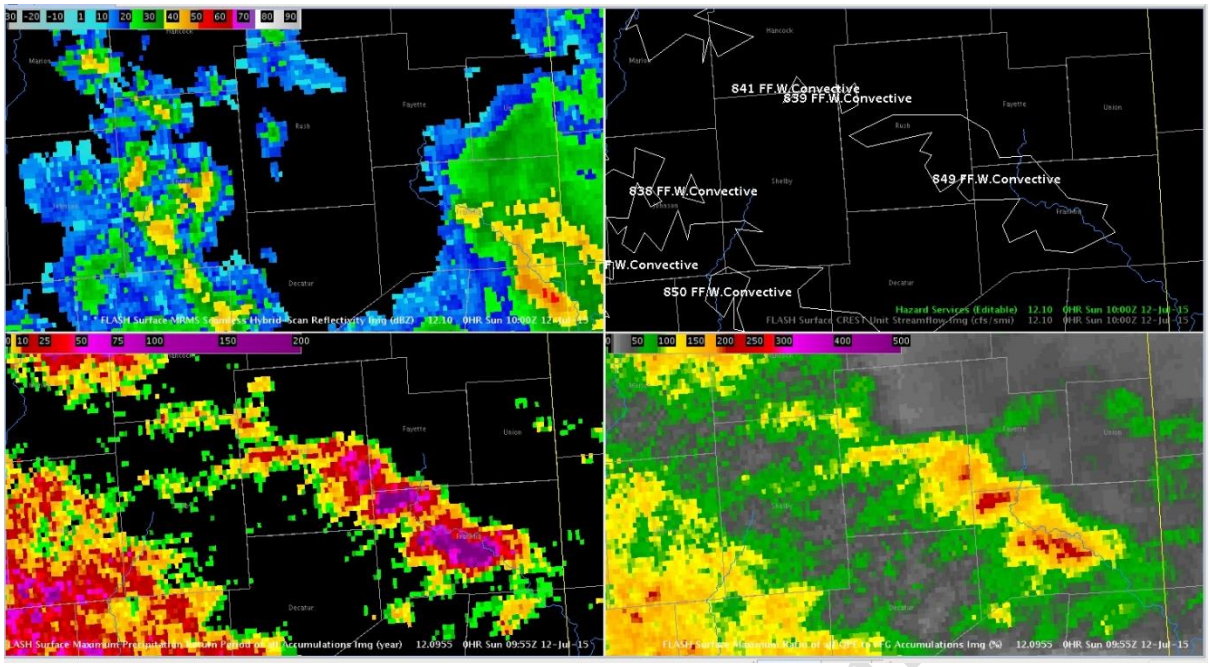

Figure 2. Demonstration of the information visualizations presented to study participants, with the automated polygons displayed in the upper-right quadrant with the underlying unit streamflow visualization hidden

2.4 Data Collection Instruments

The study utilized a Tobii TX300 eye tracking system attached beneath a 23-inch monitor. Participants interacted with the display via a keyboard and mouse. In addition to eye tracking, a set of probes assessed SA across the three theoretical levels in a method modelled on Dao et al. (2009). SAGAT-style probes assessed SA with respect to the three levels in the Endsley (1995b) model of SA, and these were presented during breaks in short scenario-based simulations.

In line with this framework, the probes assessed awareness of information in the past (the information contained within the first 1.5 hours' worth of data visualizations), the present (the final frame of data visualizations), and the future (expectations of flooding in the following two hours). Probes were presented in the form of a question about the past, present, or anticipated environmental status accompanied by a blank map of the scenario's geographic region, and participants provided responses by clicking on the maps multiple times. Some probes asked participants to recall specific values, whereas others assessed participants' abilities to synthesize information, for example, by identifying areas on a map that were at risk of flooding. Value recall-based probes were assigned a score based on response accuracy, while map-based probes were scored based on proximity of responses to the correct state. For these probes, a score was assigned on the physical proximity of the participant's response to the ground truth. An illustration of one such probe is shown in Figure 3. Participant performance

was considered per level of SA and also as a composite score based on the mean value of the scores within and across levels.

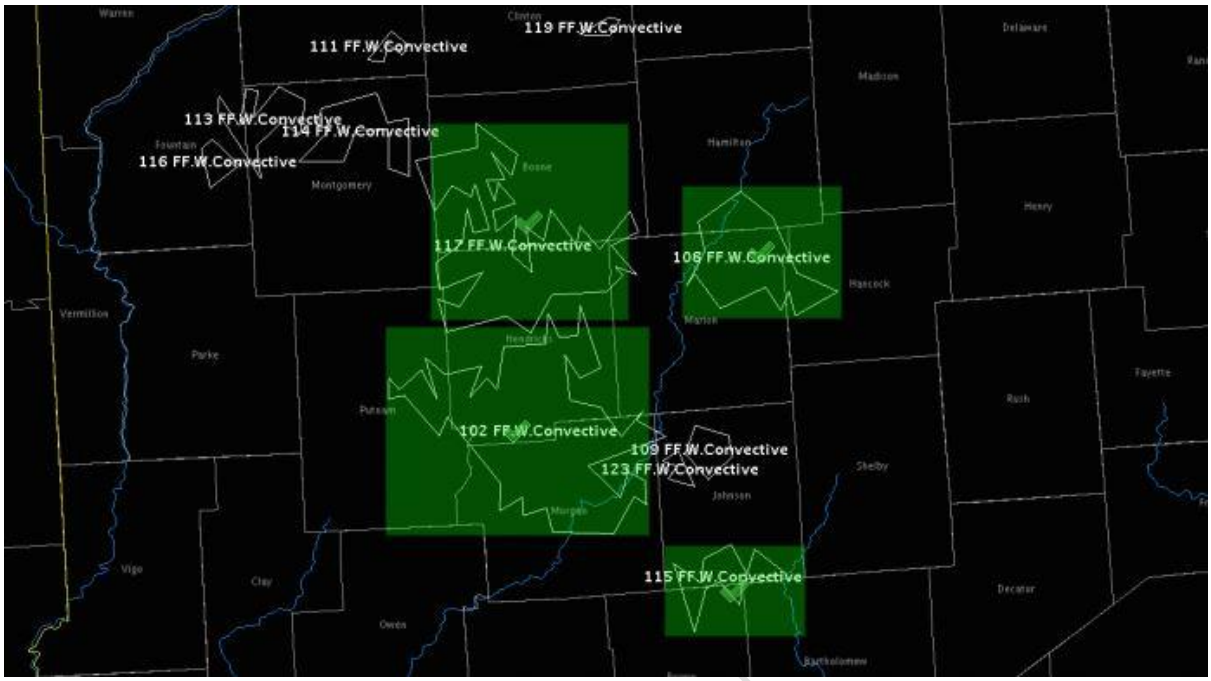

Figure 3. The study used a freeze-probe approach to assess SA, where participants responded to queries by clicking on regions of a blank map. Here, comprehension of the present environmental state was assessed, with participants responding to the statement: "Using the provided map, click which polygons highlight areas with conditions associated with flash flooding at the current time."

2.5 Procedure

Upon arrival, participants received an explanation of the study goals and signed an informed consent form. Following this, participants received training over the forecasting guidance products, which also gave participants an opportunity to ask questions about the study and the forecasting tools. When ready, participants were prompted that they had just begun a forecasting shift and a significant rainfall event was underway. In the prompt, a fictional colleague needed them to review the prior two hours of model data and radar scans in order to identify areas of high flash flooding risk. Participants completed three unique forecasting scenarios presented in a randomized order. Before each scenario began, participants read an environmental status briefing containing information about existing heavy rainfall watches, flash flood watches/warnings, and mesoscale discussions produced by the NWS Storm Prediction Center and WFOs.

Following the briefing, participants took part in one of three scenarios. Scenarios and automation conditions were presented semi-randomly. During each trial, the four-panel display showed each of the guidance products. In the automation-unavailable condition, the guidance products appeared with no alterations. However, in the automation-available condition, the polygons were overlaid on the upper-right quadrant of the display.

Although participants could view the automated polygons in the automation-available condition, they were not constantly visible. Due to technical constraints, the polygons only appeared during the timestamp that they referenced. Thus, even when available, participants were not always able to see the automation. The experimental conditions were distinguished from each other in that in one, participants could choose to use the automation, 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 eye tracker captured fixations and their duration. At breakpoints, participants responded to the SA assessment probes. In line with the modified SAGAT framework, participants indicated their responses to the map-based and value-recall probes in order to assess SA held in memory. Following this, participants either viewed the second half of the same scenario with the reverse automation condition. 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 per participant to complete.

2.6 Hypotheses

We hypothesized that automation presence would increase the number of eye fixations as well as eye fixation duration within the display panel containing the automation. We hypothesized that the greatest number of eye fixations and fixation durations would occur in the quadrant of the four-panel information display containing the automated polygons. We also assessed scanning patterns across the display, hypothesizing that the proportion of glances between AOIs would vary based on automation condition.

In relation to SA, we expected to identify a difference between automation conditions. Aligning with findings by Endsley and Kiris (1995), it was hypothesized automation presence would significantly affect Level 2. As one premise was that the attention-directing automation would guide forecaster attention to geographic areas of high risk, we also hypothesized that this would lead to higher scores on the Level 3 probes in the scenarios where automation was available. Finally, when considering eye tracking measures in conjunction with SA measures, we expected to identify a correlation between the two. The research hypotheses are presented in Table 2.

Table 2. Research hypotheses

Hypothesis	Description
H1	Situation awareness will be affected by automation condition
H2	Total fixation duration will be affected by automation condition and will vary between AOIs
H3	Fixation count will be affected by automation condition and will vary between AOIs
H4	Scanpath behavior will vary based on automation condition

H5	There will be a relationship between mean situation awareness score and at least one eye tracking measure

2.7 Data Analysis

Eye tracking dependent variables (fixation duration and fixation count) were collected and analyzed with the Tobii Studio software with four AOIs corresponding to the forecasting information visualizations in Figure 1. The data were filtered based a 70% gaze sample threshold, indicating the number of usable data points per recording. In total, this led to a sample of 68 recordings being included in the analyses. Situation awareness was measured with an accuracy-based score for the probes reflecting Levels 1, 2, and 3 SA, as well as a composite score calculated from the mean of the level scores.

The assumptions of normality and equal variance were checked for each variable. A series of t-tests, ANOVAs, and post hoc tests were used to explore the relationships between the dependent variables, automation condition, and decision aid use. Finally, a robust regression was used to explore the relationship among the eye tracking variables and SA score.

3. RESULTS

3.1 Situation Awareness Scores

Between scenarios, participants responded to probes designed to assess SA using a freeze-probe technique. In combination, findings did not support Hypothesis H1. The composite mean SA scores satisfied the assumptions of constant variance and normality, $D(104) = 0.068$, $p > 0.15$. Similarly, the normality assumption was upheld for mean scores of SA for Level 2 probes, $D(104) = 0.07$, $p > 0.15$, but not for Level 1 probes, $D(104) = 0.15$, $p < 0.01$, or Level 3 probes, $D(104) = 0.11$, $p < 0.01$. A paired, two-tailed t-test failed to identify a significant difference in composite SA score between the automation available ($\mu = 0.34$, $\sigma = 0.14$) and unavailable conditions ($\mu = 0.33$, $\sigma = 0.13$), $t(101) = -0.21$, $p = 0.83$. Individually, a Kruskal-Wallis test did not find any significant differences in performance between automation conditions in the Level 1 scores, $\chi^2(1) = 0.04$, $p > 0.15$, or Level 3 scores, $\chi^2(1) = 0.00$, $p > 0.15$; similarly, a t-test failed to identify any significant effect of automation condition on SA score for Level 2 probes, $t(100) = -0.81$, $p > 0.15$.

3.2 Eye Tracking Measures

3.2.1 Total Fixation Duration. Total fixation duration data satisfied the homogeneity of variance assumption but a Kolmogorov-Smirnov Test indicated that the data were not normally distributed, $D(269) = 0.10$, $p < 0.01$. An ANOVA found that the availability of automation did not produce significant effect, $F(1, 261) =$

2.41, $p = 0.12$. However, it did identify a significant difference between AOIs in terms of total fixation duration, $F(3, 261) = 7.92$, $p < 0.001$, as well as a significant interaction effect between automation condition and AOI, $F(3, 261) = 6.44$, $p > 0.001$, a finding in alignment with Hypothesis H2. A post hoc Tukey test indicated that among the four AOIs, participants spent the same amount of time observing the MRMS panel ($\mu = 65.21$, $\sigma = 26.89$) and the USF panel ($\mu = 52.78$, $\sigma = 39.84$), but spent significantly more time observing the MRMS than the FFG ($\mu = 43.64$, $\sigma = 28.11$) or the ARI panels ($\mu = 42.51$, $\sigma = 29.79$), but total fixation duration among the USF, FFG, and ARI were not significantly different ($p > 0.05$). A summary of the total fixation duration data by AOI and automation condition is shown in Figure 4.

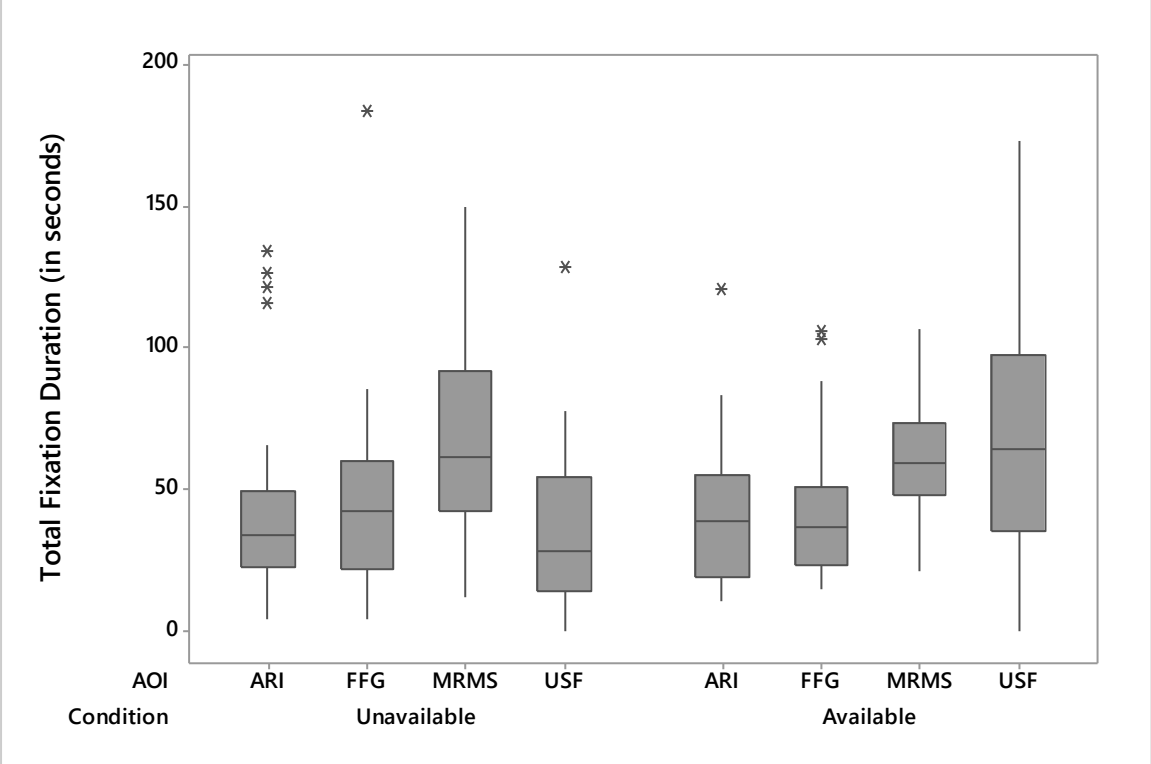

Figure 4. Total fixation duration per AOI and automation condition with central tendency and spread of data

3.2.2 Mean Number of Fixations by AOI. The mean fixation count data satisfied the equal variance assumption but was not normally distributed, $D(269) = 0.10$, $p < 0.01$. An ANOVA identified a significant interaction between the AOI and automation condition, $F(3, 261) = 6.16$, $p < 0.001$. Significant main effects were also identified with both automation condition, $F(1, 261) = 7.21$, $p < 0.01$, and AOI, $F(3, 261) = 4.28$, $p < 0.01$. A Tukey pairwise comparisons test for post hoc comparison revealed that automation availability was associated with a greater number fixations ($\mu = 148.17$, $\sigma = 84.51$) compared to the unavailable condition ($\mu = 122.62$, $\sigma = 77.27$). Post hoc comparisons also revealed that fixation count in the MRMS AOI ($\mu = 158.51$, $\sigma = 63.37$) significantly differed from the FFG panel ($\mu = 119.91$, $\sigma = 65.69$, $p < 0.05$) and the ARI panel ($\mu = 118.70$, $\sigma =$

78.21, $p < 0.05$), but not the USF panel ($\mu = 158.51$, $\sigma = 63.37$, $p > 0.05$). The remaining three AOIs did not significantly differ from each other. This may have been due to the salience of radar imagery in the MRMS AOI, or alternatively because radar imagery updated every two minutes during the simulations, which was more frequent than the other AOI types. Findings upheld Hypothesis H3, and a summary of the fixation count data by AOI and automation condition is shown in Figure 5.

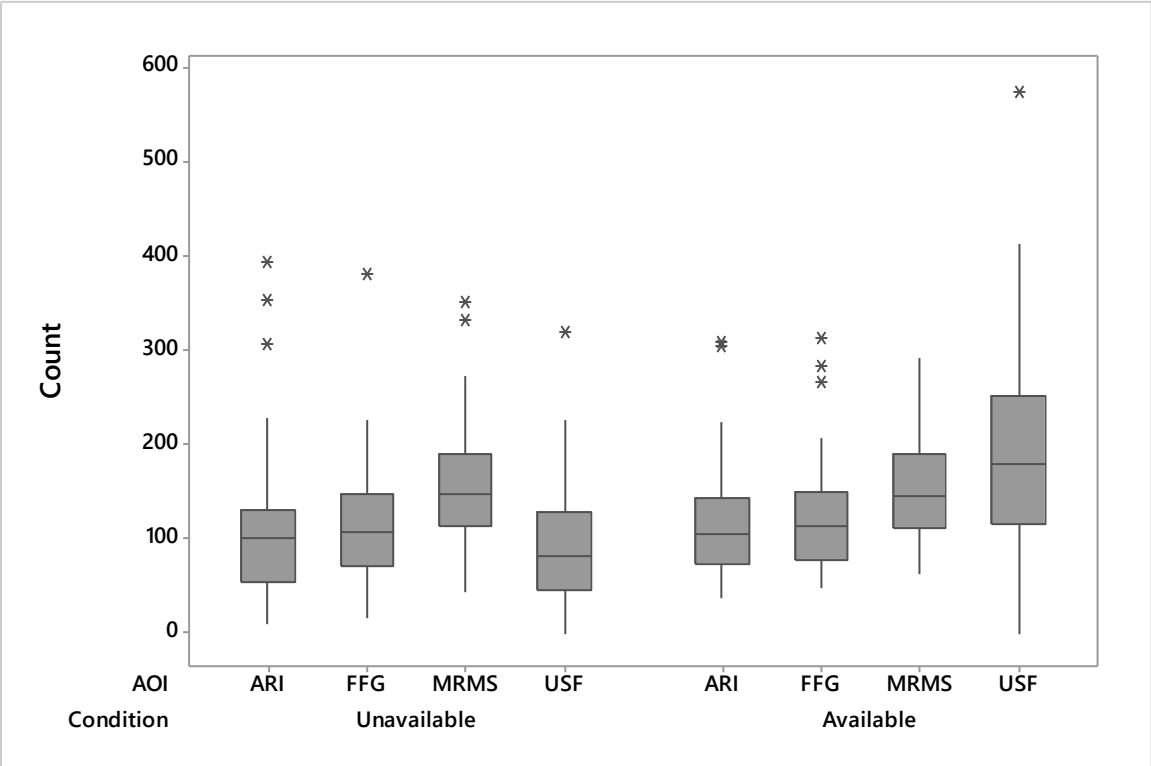

Figure 5. Number of eye fixations per AOI and automation condition with central tendency and spread of data

3.2.3 Scanpath Analysis. In addition to recording fixation data, the eye tracker also captured information about gaze patterns. Analysis of scanpaths, or the time sequenced order of fixations among AOIs, reflects scanning patterns among display elements and has been used to categorize behaviors associated with performance differences (McClung & Kang, 2016). We estimated the number of movements between the core AOIs by calculating the frequency of bidirectional exchanges. 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 6 presents a comparison between gaze movement exchanges between the automation-available and automation-unavailable conditions, shown as a percentage of all core AOI exchanges. When automation was absent, participants frequently compared the MRMS AOI with the USF and FFG AOIs. When automation was present, participants slightly changed their scanning behavior; participants had slightly fewer comparisons between the MRMS and FFG AOIs, but slightly more ARI-USF and MRMS-ARI exchanges, however, contrary

to Hypothesis H4, a t-test failed to detect a significant difference between the automation-available ($\mu = 0.17$, $\sigma = 0.036$) and –unavailable conditions ($\mu = 0.17$, $\sigma = 0.049$), $t(9) = 0.00$, $p = 1.00$.

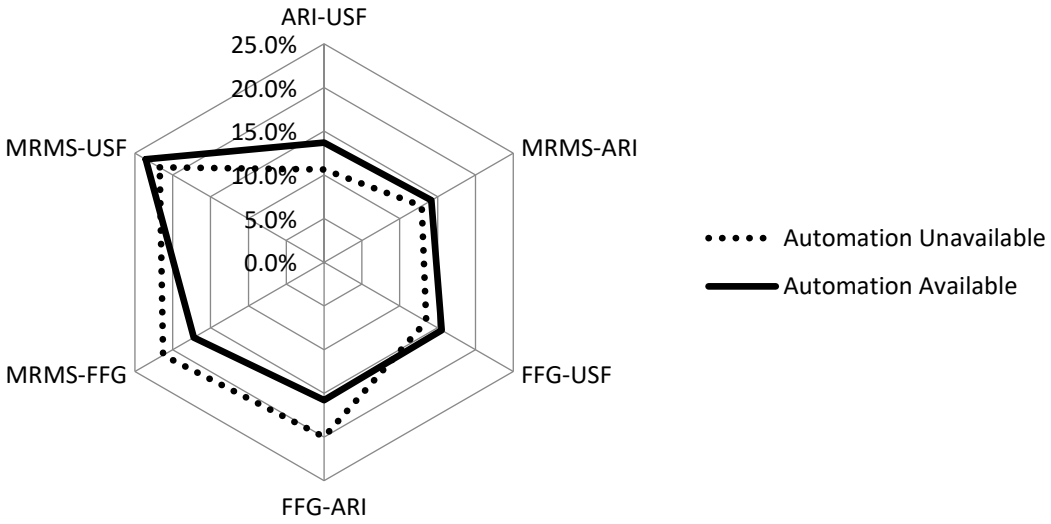

Figure 6. Inter-AOI glance movements representing scanpaths between the core AOIs with and without access to automation

3.3 Assessing the Relationship Between SA Score and Eye Tracking Measures

A multiple regression model was fit to the data to explore the relationship between eye movements and the composite SA score. In addition to the dependent variables previously discussed, the regression model incorporated an additional variable of time to first fixation within each AOI. Task duration, forecasting scenario, automation condition, and two-way interaction terms among them were also considered.

The forward selection technique identified a reduced model with an Adjusted R^2 of 0.28. Composite SA score was predicted by the total fixation duration within the ARI AOI predictor, ($\beta = 0.0019$, $p < 0.05$), number of fixations within the MRMS AOI, ($\beta = -0.00064$, $p = 0.20$), and the scenario, ($\beta_{NJ} = 0.14$, $p < 0.01$; $\beta_{WV} = -0.0018$, $p > 0.15$) ($R^2 = 0.35$; Adj. $R^2 = 0.28$). An ANOVA identified a significant effect of scenario, $F(2, 38) = 6.58$, $p < 0.01$, and ARI total fixation duration, $F(1, 38) = 4.14$, $p = 0.049$. These findings lend support to Hypothesis H5, indicating that total fixation duration and fixation count are appropriate indicators of situation awareness within forecasting tasks.

4. DISCUSSION

This study explored the relationship between situation awareness, attention-directing automation, and eye movement behavior. Findings moderately supported the hypothesis (H5) that eye tracking measures would predict performance on an accuracy-based SA assessment, a finding consistent with previous research (Moore and Gugerty 2010; van de Merwe et al. 2012). The regression analyses indicated that participant SA was predicted by total fixation duration, particularly within the ARI visualization. SA was also affected by the scenario variable, suggesting that participants' experience levels with forecasting over certain geographic regions played a significant role, or that the forecasting tasks significantly differed in demand, or perhaps a combination of both. These findings support the concept that visual attention is linked to awareness, and they show that overall SA increased as participants spent more time considering specific information sources. This may suggest that participants who spent more time fixating on the visualizations were able to develop a more comprehensive situational picture.

Against expectations, findings from the analysis of SA scores suggest that automation condition did not significantly affect participant SA (H1). This may be explained by the nature of the instructions provided during the pre-study training. Participants were instructed to treat the automation as support but not a definitive prediction. Thus, participants may have been able to remain in the loop while assessing the situation. The automation appeared to attract participants' attention; in comparison with data on the total fixation duration, the total time spent fixating within the USF AOI was the lowest in the automation unavailable condition, but the value increased with availability. This scanpath analysis revealed individual differences in forecast guidance usage during the situation assessment process. When polygons were absent, participants frequently compared the MRMS AOI with the USF and FFG AOIs. When polygons were present, participants slightly changed their scanning behavior; participants had fewer comparisons between the MRMS and FFG AOIs, but slightly more ARI-USF and MRMS-ARI exchanges, however, contrary to Hypothesis H4, these patterns were not significantly different between automation conditions. It is possible that some of these minor differences were due to variations in individual expertise, familiarity with flash flood forecasting guidance products, or level of understanding with regard to the automated polygons, but further investigation would be needed to clarify this.

Closer inspection of the eye movement data confirmed several but not all of the hypotheses related to the automation condition. The hypothesis that participants would fixate more frequently within the USF AOI when the automation was present was confirmed (H3), indicating that participants did attend to the automated forecast polygons when they were visible. Likewise, findings also partially confirmed the hypothesis that fixation duration in the USF AOI would increase with automation availability (H2). This phenomenon was observed in the

proportional fixation duration data, but not the absolute duration data. In comparison, participants spent the most time fixating within the most familiar product, the radar scans. In addition to being commonly used in operations, radar imagery received updates more frequently than the hydrologic models did, meaning that the visual stimuli changed more frequently. This may have attracted the eyes and motivated participants to reassess the panel more often than the other AOIs. Participants had varying levels of experience with the other three guidance products with the least familiar being the USF visualization.

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 current automation 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 hypothesize that SA may improve as operators gain experience with automation. Furthermore, while Dao et al. (2009) did not detect any significant difference in SA accuracy between automation levels in an air traffic control task, they did find that SA, as measured by response time, did correlate to automation level. In the current work, a preliminary investigation did not reflect a significant correlation between response times and the current results, but further research would be needed to explore this further.

While the regression analyses demonstrated eye tracking's viability as a method for assessing SA, it also provided insight at a qualitative level into how forecasters developed and maintained SA. The analyses showed that in addition to consulting individual products, forecasters routinely compared between all four products. The comparative nature of situation assessment has been identified as an integral part of weather forecasting (Kirschenbaum 2004; Pliske et al. 2004; Trafton and Hoffman 2007). Eye tracking supports this line of research through scanpath analysis, which has been used to describe information scanning during forecast decision making (Wilson et al., 2016) and have been categorized for human performance analysis (Kang & Landry, 2014; Kang & Landry, 2015; McClung & Kang, 2016). In the current investigation, relatively long fixation durations and high fixation counts suggest that participants not only used the familiar radar imagery to establish a situational understanding but they also used it in comparison tasks as a benchmark. Feedback received during the debriefing phase indicated that some participants used the familiar radar product to calibrate their mental models with the less-familiar guidance products. Furthermore, in addition monitoring and comparison activities, SA in forecasting is affected by a collection of cognitive and system factors. For example, from a system design perspective, visualization algorithms have been shown to affect SA in forecasting; in a study where participants judged ARI

visualizations depicting flash flood conditions, Argyle et al. (2017) observed a difference in threat detection abilities between two alternative visualization algorithms. Forecaster characteristics also play a role in SA; Gugerty (2011) suggested that attention allocation capacity could impact an operator's ability to develop SA. While not the main objective of this study, attention allocation can also be assessed with eye tracking, and it incorporating this into models of human decision making could offer benefits to performance monitoring while enhancing understanding of cognitive phenomena.

In addition to providing insight into the use of eye tracking to describe forecaster information usage, this study has several practical implications, first, in relation to eye tracking as a less-invasive mechanism for assessing SA, and second, in relation to the use of eye tracking assessment as a means for supporting the development of future forecasting visualization systems. Findings from the regression analysis supported the hypothesis that SA would be predicted via ocular measures (H5). The findings aligned with those of Moore and Gugerty (2010), who observed that the proportion of fixation time within an AOI was the most significant predictor of SA in an air traffic control task. In the current work, the analysis indicated that total fixation duration, fixation count, and task scenario acted as significant predictors of SA level in this context. Performance-based SA measures, such as SAGAT, are relatively easy to implement and can provide insight into each of the three levels of SA (Endsley 1995a), but they have been criticized for being disruptive to participant workflow (Salmon et al. 2008). While previous studies have used eye fixation data to group participants into high- versus low-SA levels (Bhavsar et al. 2017; Kiran et al. 2018), this study demonstrates that fixation data may be able to predict SAGAT-based measures of SA, thereby opening the door to more objective and less intrusive assessments of human cognition.

A second practical implication of this work relates to the development of future forecasting decision aids. The regression analysis indicated that total fixation duration (ARI AOI) and fixation count (MRMS AOI) predicted overall SA as estimated by the mean score from the probes. While a link to radar imagery was not surprising, the significance of fixation duration within the ARI visualization was more so, and this introduced a new question regarding the utility of other in-development flash flood prediction models. Although radar scans and FFG-based guidance products are available in operational settings, the Average Recurrence Interval (ARI) and CREST Unit Streamflow (USF) products were in-development at the time of this work and not available for use in formal work display systems. The inclusion of the ARI fixation duration term in the regression model suggests that it played a role in developing awareness of the unfolding flash flooding situation. In future work, a similar eye tracking-based methodology could be applied in further investigations of the influence of novel

decision aids on forecaster situation awareness, where findings could inform the user-centered design of information visualization systems.

There were several limitations that would potentially limit generalizability of the outcomes. First, participants had varying levels of familiarity with the geographic regions used in the scenarios. For example, one participant had extensive experience forecasting in the Indiana-based scenario, and so he was very familiar with county names and river structures presented in the Indiana scenarios. Others, however, expressed discomfort with being able to assess risk accurately in regions with which they were less familiar. As past research into forecasting, experience, and SA has shown, experience and mental models are closely linked with SA, and such a lack of experience may be a contributing factor to diminished SA during this experiment.

In terms of technical limitations, at the time of the study, forecasting automation was within conceptual stages of development, requiring the tool used within the study to be an abstraction of an envisioned future product. The algorithm used in the present study was based on a threshold from the QPE Ratio product, which was chosen due to FFG's familiarity to most forecasters across the United States. While the automation highlighted the regions of interest that were visible in the QPE-to-FFG ratio map, it did provide a way to directly compare the two products in a way that overlaying the two maps on top of each other could not (overlaying tended to lead to maps that were hard to interpret when color tables conflicted). Forecasters often engage in comparison activities and so this may be an effective way to facilitate comparison and mental model building (Kirschenbaum, 2004; Trafton & Hoffman, 2007).

Finally, few participants were familiar with the concept of forecasting automation, which likely impacted general understanding in terms of intended usage, despite training. Nevertheless, the eye tracking measures showed that participants did consult the AOI containing automation in those scenarios where it was available. As forecasting automation takes 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 automation, 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.

5. CONCLUSIONS

This study supports the hypothesis that eye tracking can provide a minimally invasive means of estimating SA in dynamic, sociotechnical systems. Future work should investigate how weather forecasters develop and maintain SA, particularly during tasks involving decision support. In addition to identifying a link

between SA and eye tracking measures, the analysis revealed behavioral differences when participants had access to automation. While the automation did not appear to distract from the main forecasting task, participants had relatively low SA. In future work, employing a more naturalistic research approach with eye tracking may facilitate real-time data capture from forecasters in operational settings in a relatively non-intrusive manner, and could continue to progress the technology towards an in-situ application. Eye tracking has potential for use in operational forecasting environments, and it may be an effective tool for identifying forecasters' training needs or for real-time assessment of human performance.

The greatest contribution of this work is the identification of the significant relationship among eye fixation data and probe-based measures of SA in weather forecasting. Previous studies have used eye tracking as an SA assessment method, but research has frequently focused on aviation, air traffic control, and driving. The study provided evidence into the impact of an attention-directing automation mechanism on forecaster SA, and from a methodological perspective, also showed that several eye tracking metrics can be used, to varying degrees, as predictors of SA.

7. ACKNOWLEDGEMENTS

Funding for this research was provided by the Disaster Relief Appropriations Act of 2013 (P.L. 113-2), which provided support to the Cooperative Institute for Mesoscale Meteorological Studies at the University of Oklahoma under Grant NA14OAR4830100. The authors would like to thank Gabe Garfield and the staff of the Hazardous Weather Testbed for providing an environment to conduct this study. We also appreciate the support provided by Zachary Flamig, Katie Wilson, Chen Ling, Pamela Heinselman, Scott Gronlund, and Theodore Trafalis during the development of this work.

8. REFERENCES

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.

Argyle, E. M., Gourley, J. J., Ling, C., Shehab, R. L., & Kang, Z. (2017). Effects of Display Design on Signal Detection in Flash Flood Forecasting. *International Journal of Human-Computer Studies, 99*, 48-56.

Bhavsar, P., Srinivasan, B., Srinivasan, R. (2017). Quantifying situation awareness of control room operators using eye-gaze behavior. *Computers & Chemical Engineering*, *106*, 191-201.

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.

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.

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. Human Factors, 37(1), 65-84.

Endsley, M. R. (1995b). 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. & Hoffman, R. R. (2002). The Sacagawea Principle. Intelligent Systems, IEEE, 17(6), 80-85.

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.

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.

Gourley, J. J., Flamig, Z. L., Vergara, H., Kirstetter, P.-E., III, R. A. C., Argyle, E., . . . Howard, K. W. (2016). The Flooded Locations And Simulated Hydrographs (FLASH) project: improving the tools for flash flood monitoring and prediction across the United States. *Bulletin of the American Meteorological Society, 98*(2), 361-372.

Heine, T., Lenis, G., Reichensperger, P., Beran, T., Doessel, O., & Deml, B. (2017). Electrocardiographic features for the measurement of drivers' mental workload. *Applied Ergonomics, 61*, 31-43.

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.

Ikuma, L.H., Harvey, C., Taylor, C.F., Handal, C. (2014). A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. *Journal of Loss Prevention in the Process Industries, 32*, 454-465.

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.

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.

Kang, Z. & Landry, S. J. (2014). Using scanpaths as a learning method for a conflict detection task of multiple target tracking. *Human Factors, 56*(6), 1150-1162.

Kang, Z. & Landry, S. J. (2015). An eye movement analysis algorithm for a multielement target tracking task: Maximum transition-based agglomerative hierarchical clustering. *IEEE Transactions on Human-Machine Systems, 45*(1), 13-24.

Kiran, R., Salehi, S., Jeon, J., & Kang, Z. (2018). Real-Time Eye-Tracking System to Evaluate and Enhance Situation Awareness and Process Safety in Drilling Operations. Paper presented at the IADC/SPE Drilling Conference and Exhibition, Fort Worth, Texas, USA, 2018/3/6/

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.

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

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.

Marinescu, A. C., Sharples, S., Ritchie, A. C., López, T. S., McDowell, M., & Morvan, H. P. (2018). Physiological Parameter Response to Variation of Mental Workload. *Human Factors, 60*(1), 31-56.

McClung, S. N. & Kang, Z. (2016). Characterization of Visual Scanning Patterns in Air Traffic Control. *Computational Intelligence and Neuroscience, 2016*, 17.

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.

Pierce, R. S. (2012). The Effect of SPAM Administration During a Dynamic Simulation. *Human Factors, 54*(5), 838-848.

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.

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*, 297-323.

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.

Trafton, J. G. & Hoffman, R. (2007). Computer-aided visualization in meteorology. In R. Hoffman (Ed.), *Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making* (pp. 337-357). New York, USA: Lawrence Erlbaum Associates.

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

Wilson, K. A., Heinselman, P. L., & Kang, Z. (2016). Exploring Applications of Eye-Tracking in Operational Meteorology Research. *Bulletin of the American Meteorological Society, 97*(11), 2019-2025.

Wilson, K. A., Heinselman, P. L., Kuster, C., Kingfield, D. M., & Kang, Z. (2017). Forecaster Performance and Workload: Does Radar Update Time Matter? *Weather and Forecasting, 32*(1), 253-274.

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.

Yu, C. S., Wang, E. M., Li, W. C., Braithwaite, G., & Greaves, M. (2016). Pilots' Visual Scan Patterns and Attention Distribution During the Pursuit of a Dynamic Target. *Aerospace Medicine and Human Performance, 87*(1), 40-47.