Comparing Forecaster Eye Movements during the Warning Decision Process

KATIE A. WILSON

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

PAMELA L. HEINSELMAN

NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

ZIHO KANG

School of Industrial and Systems Engineering, University of Oklahoma, Norman, Oklahoma

(Manuscript received 17 August 2017, in final form 26 January 2018)

ABSTRACT

An eye-tracking experiment was conducted to examine whether differences in forecasters' eye movements provide further insight into how radar update speed impacts their warning decision process. In doing so, this study also demonstrates the applications of a new research method for observing how National Weather Service forecasters distribute their attention across a radar display and warning interface. In addition to observing forecasters' eye movements during this experiment, video data and retrospective recalls were collected. These qualitative data were used to provide an explanation for differences observed in forecasters' eye movements. Eye movement differences were analyzed with respect to fixation measures (i.e., count and duration) and scanpath dimensions (i.e., vector, direction, length, position, and duration). These analyses were completed for four stages of the warning decision process: the first 5 min of the case, 2 min prior to warning decisions, the warning issuance process, and warning updates. While radar update speed did not impact forecasters' fixation measures during these four stages, comparisons of scanpath dimensions revealed differences in their eye movements. Video footage and retrospective recall data illustrated how forecasters' interactions with the radar display and warning interface, encounters with technological challenges, and varying approaches to similar tasks resulted in statistically significantly (p value < 0.05) lower scanpath similarity scores. The findings of this study support the combined use of eye-tracking and qualitative research methods for detecting and understanding individual differences in forecasters' eye movements. Future applications of these methods in operational meteorology research have potential to aid usability studies and improve human–computer interactions for forecasters.

1. Introduction

Understanding the forecaster warning decision process is a complex task that has been at the forefront of the Phased Array Radar Innovative Sensing Experiment (PARISE) since 2010. Learning about the potential impacts of rapidly updating phased-array radar (PAR) data on forecasters' warning decision processes requires not only an assessment of performance, but an in-depth analysis of how forecasters acquire, make sense of, and use information to provide the best possible warnings (Heinselman et al. 2012, 2015; Bowden et al. 2015; Bowden and Heinselman 2016). Other studies within the NOAA Hazardous Weather Testbed have evaluated forecasters' use of new data and displays using qualitative methods including observations, surveys, discussion, interviews, and blog posts (e.g., Goodman et al. 2012; Calhoun et al. 2014; Smith et al. 2014; Karstens et al. 2015; Smith et al. 2016). Furthermore, in PARISE, cognitive task analysis methods have been applied to obtain detailed insight into what forecasters see, think, and do when presented with radar data of different update speeds (e.g., Heinselman et al. 2015; Bowden and Heinselman 2016). Referred to as the recent case walkthrough (Ericsson and Simon 1993; Hoffman 2005), this method required forecasters to retrospectively recall their thought processes as they watched a playback video of their onscreen activity that

Corresponding author: Katie Wilson, katie.wilson@noaa.gov

DOI: 10.1175/WAF-D-17-0119.1

was recorded during simulated warning operations. As forecasters recalled their thought processes step by step, they were also asked probing questions that focused on times when warning decisions were made.

Much has been learned from retrospective recall data about how faster radar updates can impact forecasters' warning decision processes during different types of severe weather scenarios. However, these data have also brought to light how complex forecasters' warning decision processes can be and that the use of qualitative methods alone may not always accurately capture the intricate activity occurring within a forecaster's mind. Therefore, a method that could provide additional measures of forecasters' cognition was sought, where together, this additional data would be analyzed with the recent case walkthrough to inform the interpretation of one another.

Given that studies have shown that what we are looking at can be informative of what we are attending to (Just and Carpenter 1976; 1980), eye-tracking research methods were explored as a means to better trace forecasters' cognition when they are viewing radar data. The use of eye-tracking methods was first applied in reading studies to learn about language processing, and fixations and saccades were identified as two types of eye movements (Huey 1908; Rayner 1998; Duchowski 2002; Henderson and Ferreira 2004). Fixations describe times when the eye is relatively still, and saccades describe the very fast eye movements that occur between fixations. Applications of eye tracking to study other human cognition were also demonstrated in free-viewing tasks to prove that the location of fixations was not random. Rather, fixations occurred more frequently in the most semantically and visually rich regions of an image (e.g., Buswell 1935; Yarbus 1967). This observation was important because it provided evidence that eye movements are an important representation of attention.

More recently, eye tracking has been used in a variety of visual search tasks. For example, studies that have focused on web design and marketing have learned much about how the general population attend to and gather information from computer displays, advertisements, and package designs (e.g., Djamasbi et al. 2010; Hervet et al. 2011; Clement et al. 2013; Gidlöf et al. 2013; Romano Bergstrom et al. 2013; Wang et al. 2014). Additionally, eye tracking has been used to better understand the cognitive processes of professionals who make life-saving decisions. Within the medical field, many studies have examined the visual search behavior of radiologists tasked with detecting abnormalities and diagnosing medical conditions (e.g., Wood et al. 2013; Manning et al. 2004; Giovinco et al. 2015; Bertram et al. 2016; Al-Moteri et al. 2017). In aviation research, eye tracking has been used to study the eye movements of pilots in the cockpit and air traffic controllers on the ground (e.g., Hauland 2008; Sullivan et al. 2011; Van de Merwe et al. 2012; Kang and Landry 2014, 2015; Yu et al. 2016). A common interest in these medical and aviation studies is how visual scanning patterns compare between professionals with different degrees of experience and whether observed differences can inform training to improve performance. Like professionals working in aviation and medicine, meteorologists are also presented with imagery to analyze. They must distribute their attention according to the task at hand, assess how targets (i.e., storm features) are evolving in time and space, and perceive and extract information considered important to their decision-making. Eye-tracking studies from the aviation and medical communities are therefore highly relevant to our quest to better understand meteorologists as expert decision-makers.

Despite the growing popularity of eye-tracking methods in other research domains, it has been applied in only a handful of meteorology studies. For example, eye movement data have been used to analyze what impact a weathercaster's gesturing would have on viewers during a televised weather forecast (Drost et al. 2015), to assess the impact of legend color and content on participants' abilities to correctly interpret hurricane storm surge graphics (Sherman-Morris et al. 2015), and to study how U.S. Naval and Marine Corps weather forecasters extract information from meteorological visualizations (Trafton et al. 2002). In an exploratory sense, Wilson et al. (2016) assessed the feasibility of eye-tracking research methods for building on the current understanding of forecasters' warning decision processes. Without previous examples of NWS forecasters' eye movement data, a simple question was whether a forecaster's eye movement data would make sense and be representative of their experienced cognition. In this short study, Wilson et al. (2016) collected a single NWS forecaster's eye movement data as he interrogated radar data during simulated warning operations. This participant's retrospective recall was also collected following the simulated event. Comparing trends in these eye movement data to the participant's retrospective recall, this study concluded that the eye movement data were representative of that forecaster's warning decision process and his reporting of important events during the simulation (e.g., a change in the expected weather threat and his subsequent redistribution of attention) (Wilson et al. 2016). We therefore anticipated that other forecasters' eye movements would also be representative of their own warning decision processes, as verified through an analysis of their retrospective recall narratives.

The findings from Wilson et al.'s (2016) study supported the use of eye tracking as a method for observing the visual attention of a forecaster. These findings motivated a larger-scale experiment that we present in this paper. Of particular interest was how forecasters' eye movements compare during use of different radar update speeds. Given that previous studies have shown radar update speed to affect forecasters' performance and overall situational awareness (e.g., Heinselman et al. 2015; Bowden and Heinselman 2016; Wilson et al. 2017b), we were specifically interested in whether differences in forecasters' related warning decision processes would be evident in their eye movements. This study explored what, if any, differences existed between forecasters' eye movements while they worked a single weather event with either 1- or 5-min radar updates. Forecasters' corresponding retrospective recalls and computer video data were then used to make sense of their eye movements within the broader context of the warning decision process. As is apparent in the results section below, we did not find differences in forecasters' eye movements during this weather event to be tied to the independent variable of radar update speed, but rather due to a variety of other factors. Therefore, although the discussions in this paper do not focus on forecasters' eye movements as a result of radar update speed as initially intended, the shared findings contribute to our current limited knowledge of how eye tracking can be used alongside qualitative research methods to learn more about the human component of weather forecasting.

2. Method

a. Experiment design

Over six weeks in the summer of 2015, 30 NWS forecasters from 25 Weather Forecast Offices visited the NOAA Hazardous Weather Testbed in Norman, Oklahoma, to participate in the 2015 PARISE. The largest of its kind, this most recent PARISE comprised three studies: the traditional experiment (Wilson et al. 2017b), the eye-tracking experiment, and the focus group (Wilson et al. 2017a). This paper presents results related to the eye-tracking experiment only. In this eye-tracking experiment, forecasters worked a 1-h-long event independently in simulated real time. Forecasters were randomly assigned to either a control (5-min radar updates) or an experimental (1-min radar updates) group. The control group's radar update speed was chosen to correspond approximately to the temporal resolution of radar data that NWS forecasters currently have available during warning operations. Both groups had 15 participants, all of whom were qualified to issue weather warnings. During the case, forecasters were provided with reflectivity and velocity base products only and were able to display these data using all tilts in the Warning Decision Support System–Integrated Information (WDSS-II) software (Lakshmanan et al. 2007). Given that forecasters were not familiar with WDSS-II, training on how to setup and navigate through the radar data and issue warning products was provided. A warning generation (WarnGen) tool similar to what forecasters use in operations was developed for WDSS-II, and all issued warning products were recorded in an electronic database. As in previous PARISE studies, a prebriefing video lasting several minutes was provided prior to working the case to allow forecasters to form expectations for how the weather event may unfold. This video described the environmental conditions associated with the upcoming weather event and showed prior radar and satellite data leading up to the case start time. Once forecasters had watched the prebriefing video, they were asked to work the weather event with their normal approach to data interrogation and to make warning decisions if considered necessary.

b. Weather scenario

The chosen weather scenario included a multicell severe hail and wind event that occurred during 2230–2330 UTC 8 July 2014. In addition to meeting a suitability criteria for experimental testing (i.e., uninterrupted radar observations for a sufficient duration), discussions with the NWS forecaster who worked the event in real time influenced the case selection. After viewing 1-min PAR updates of this event, the forecaster reported being able to better track cycling trends in rapid core development aloft compared to when he had used the 5.1-min WSR-88D volume updates (C. Kuster 2017, personal communication). We therefore anticipated that this case would present forecasters with an opportunity to demonstrate differences in their warning decision processes when using 1- or 5-min PAR volumetric updates.

In this scenario, the 90° PAR sector scanned toward the southeast and encompassed two areas of storms (Fig. 1). The storm in the western portion of the sector is referred to as the McClain storm, and the storm in the eastern portion of the sector is referred to as the Pontotoc storm. The discreet nature of these storms further encouraged the selection of this case, since it allowed for a clear-cut analysis of how attention was distributed between the storms. According to the official NWS *Storm Data* records (https://verification.nws.noaa.gov), only the McClain storm was associated with severe hail (at 2304 and 2328 UTC) and wind (at 2325 UTC) reports. Although the Pontotoc storm was not associated

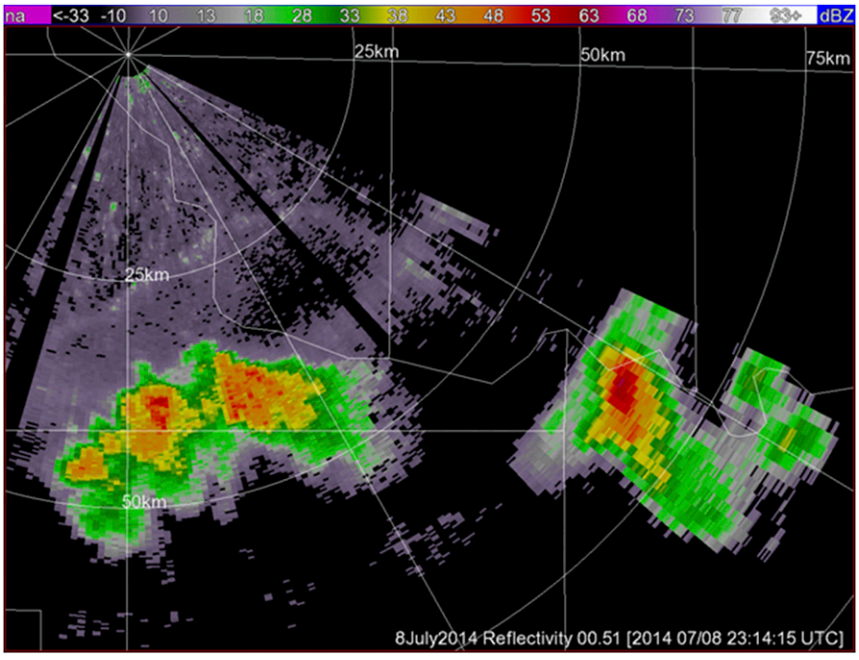

FIG. 1. Snapshot of the 0.5° reflectivity data for the (left) McClain and (right) Pontotoc storms at 2314 UTC 8 Jul 2014.

with severe weather reports in *Storm Data*, this storm presented more impressive characteristics in radar data and had higher values of maximum estimated size of hail (Witt et al. 1998) than the McClain storm. Therefore, it is possible that the Pontotoc storm also produced severe weather but that it was not observed nor reported.

c. Equipment and data collection

Both head-mounted and remote-based eye-tracking systems are available for research use (Goldberg and Wichansky 2003). Head-mounted systems are worn, and while some devices are bulky and obstruct a person's view, others have simpler designs that resemble a pair of glasses. While these systems allow for large head movements during data collection and provide a way for observing eye movements in natural settings, they can be uncomfortable and do not allow participants to engage in tasks without being fully aware that their eyes are being observed. Remote-based systems are typically positioned beneath or within a computer monitor that display the task at hand. Advancements in this technology now allow for some head movement, which makes this method less invasive because the use of chin rests to stabilize participants during observation is not necessary. An additional advantage to remote-based eye-tracking systems is that the observed visual scene stays within the same boundaries for the duration of an experiment, making data analysis much more straightforward.

Given the stationary nature of this experiment, the Tobii TX300 eye-tracking system was used to collect forecasters' eye gaze data. This remote video-based system uses infrared illumination to track pupil and corneal reflection. More specifically, dark-pupil eye-tracking methods were used, such that the infrared illumination was positioned away from the optical axis, causing the pupil to appear darker than the iris. The video camera in the eye-tracking system acquired an image of the eye at a sampling rate of 300 Hz. Through the use of image processing algorithms, the dark pupil and corneal reflection were identified, and geometrical calculations, as well as information from each forecaster's calibration, were used to map the point of vision to x and y coordinates on the computer screen. The gaze accuracy, referring to the possible angular distance error from the actual to the observed point of gaze, is 0.4° (Tobii Technology 2014). This accuracy corresponds to a 4.4-mm possible error in gaze location on the computer screen. The gaze precision of this system, referring to the spatial angular variation between gaze samples, is 0.07° (Tobii Technology 2014).

The calibration procedure each forecaster completed prior to beginning the case required them to watch the computer screen and follow a series of dots as they appeared. To ensure calibration was completed successfully, we also asked each forecaster to spend a short time browsing a web page. We used this sample of eye gaze data to ensure that the eye tracker captured the

point of vision accurately. Once calibration was completed, the Tobii TX300 was used to collect each forecaster's eye gaze data for the full duration of the weather scenario. The remote eye-tracking system was positioned beneath the computer screen, and although forecasters had to remain relatively still while working the case, some gentle head movements were allowed.

At the end of the case, the collected eye gaze dataset was checked to ensure that the gaze sample was sufficient. The gaze sample is a measure that indicates the proportion of samples that were collected successfully, is given as a percentage, and is considered acceptable for values of at least 75% (Hvelplund 2014). Data loss resulting in gaze samples below this value can occur because of difficulty in detecting the pupil and corneal reflection, possibly as a result of a person's eye color, eye shape, use of eyewear, or use of makeup. Furthermore, visual inspection of the overlaid eye gaze data on the screen recording was important for ensuring sufficient accuracy and precision of forecasters' eye gaze data. Based on these data quality checks, six datasets were removed from the analysis, and the results presented in this paper are therefore based on eye gaze data belonging to 12 participants in each group. The control group participants' experience ranged from 1 to 25 years (mean = 9.1, standard deviation = 7.1), and the experimental group participants' experience ranged from 2 to 27 years (mean = 11.8, standard deviation = 7.3).

Each forecaster also provided a retrospective recall of their warning decision process using the recent case walkthrough method (Hoffman 2005). As described in the introduction, this method was used extensively in the 2012 and 2013 PARISE studies (Heinselman et al. 2015; Bowden and Heinselman 2016). We asked forecasters to verbalize their thought processes while watching a playback video of their onscreen activity. The following instructions were provided as follows: We will now record key aspects of your warning decision process. Key aspects include what you were: seeing (e.g., Noticing reflectivity values of 60 dBZ up to 20 kft), thinking (e.g., The strong BWER suggests to me that there is a strong updraft), and doing (e.g., I am going up and down in time to see how deep the mesocyclone is). Concurrently, the assisting researcher typed the forecasters' verbalizations into a timeline. Probing questions were asked to gather further insight into why forecasters made warning decisions, including the following: 1) What was your warning decision? 2) Why did you make this warning decision? 3) How did the temporal resolution of the radar data impact this warning decision? Once forecasters had completed this retrospective recall for the weather event, they self-checked the timeline to ensure that their verbalizations were recorded accurately.

d. Data analysis

1) FIXATION IDENTIFICATION

Eye fixations and gaze patterns are of interest because they can be informative of a person's reasoning (Henderson and Ferreira 2004). To identify fixation events, the raw eye gaze data were parsed through a velocity-threshold identification (I-VT) algorithm using the Tobii Studio 3.3.0 software (Komogortsev et al. 2010; Olsen 2012). This algorithm's output lists the timestamp, duration, and x and y positions for each fixation. The x and y positions are based on a pixel grid system of the computer screen (1920 × 1080 pixels). Eye gaze velocity is described in terms of visual angle ($^\circ s^{-1}$) and is calculated as the angle between two samples divided by their separation in time. To reduce measurement noise effects, angular velocity is calculated for a 20-ms window that is centered on the sample of interest (Olsen 2012). The timestamp and position information of the first and last samples of the window determine the angular velocity of the center sample. Samples having an angular velocity below the default velocity threshold parameter ($30^\circ s^{-1}$) are classified as fixations (Olsen 2012; Bojko 2013). Adjacent fixations may either remain separate or be merged into a single longer fixation depending on the time and visual angle between them. The "max time between fixations" parameter is given as 75 ms (allowing for blink events), and the "max angle between fixations" parameter is set at 0.5° (Komogortsev et al. 2010). Two adjacent fixations become merged if the time and angle between them is less than or equal to these parameter values. Finally, a minimum fixation duration parameter of 60 ms was chosen. This minimum fixation duration was chosen because in addition to viewing radar imagery, forecasters were required to read text while creating and issuing warnings as well as when sampling storms and processing the readout information (e.g., "60 dBZ, 30 000 ftAG"). Fixations during reading studies have shown to last between 60 and 500 ms (Liversedge and Findlay 2000). All fixations with durations shorter than 60 ms were discarded.

2) AREAS OF INTEREST AND FIXATION MEASURES

In addition to identifying eye fixation events, the Tobii Studio 3.3.0 software was used to manually draw areas of interest (AOIs) that define separate spaces on the computer display (Holmqvist et al. 2011; Bojko 2013). The AOI boundaries were agreed upon by all authors and applied consistently to all forecasters' eye gaze data. These AOIs represent different semantic content, including reflectivity data, velocity data, control icons,

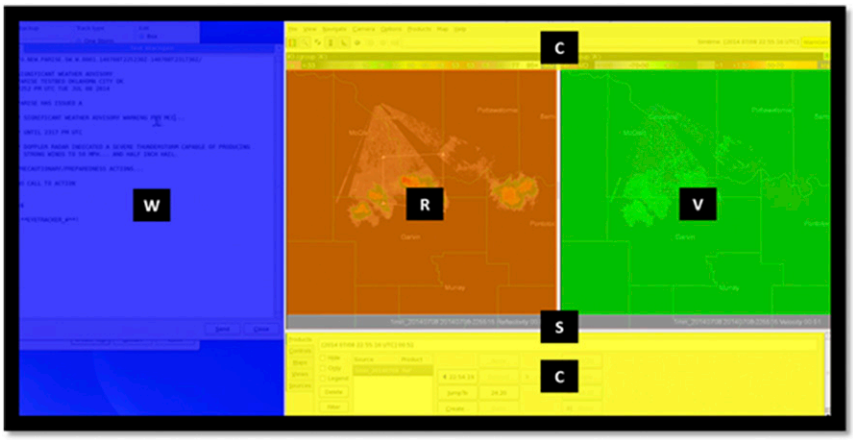

FIG. 2. Areas of interest are identified for the reflectivity data (R, orange), velocity data (V, green), control icons (C, yellow), radar scan information (S, gray), and the WarnGen interface (W, blue). Note that the WarnGen interface appeared only when the forecasters elected to use it.

radar scan information, and the WarnGen interface (Fig. 2). The two control icon areas were combined in the analysis. All identified fixations were tagged with the AOI in which they occurred. The AOI-based labeling of fixations is useful for comparing forecasters' sense-making within these spaces for different portions of the warning decision process. While many different types of fixation measures exist, two of the most commonly used measures are count and duration (Jacob and Karn 2003). We can assess within each AOI how many times forecasters fixated (count) and on average how long those fixations lasted (duration). Higher fixation counts within an AOI indicate that the information was more noticeable or important to the participant, while an AOI associated with longer fixation durations indicates that the information was either more difficult to extract or more engaging to the participant (Poole and Ball 2006; Bojko 2013). However, we recognize the limitations that forecasters' eye movements do not represent all aspects of their cognition and that the eye can drift during thought. These known limitations were a major driver for applying a mixed-methods approach during this study.

3) SCANPATH COMPARISONS

Fixation measures are useful for obtaining an overall impression of how visual attention is distributed across AOIs for a given time frame. These measures can be used to indicate whether the control and experimental groups visually attended to the different AOIs in a similar manner or not. However, these bulk measures are not good at representing how attention is distributed over time. Additionally, the spatial resolution of fixations is reduced to the size of the AOIs, meaning that the spatial distribution of fixations within an AOI is not represented either. How fixations change in time and space is an important consideration if forecasters' underlying cognitive processes during this simulation are to be understood (Noton and Stark 1971; Holmqvist et al. 2011). Therefore, in addition to average AOI fixation measures, the sequence of fixations in time and space was examined with AOI boundaries removed.

In this study, it was essential to maintain both the temporal ordering of the entire sequence of fixations and the spatial resolution for adequate representation of scanpath shape in similarity calculations (Jarodzka et al. 2010). Therefore, the MultiMatch method was selected for the scanpath comparison analysis. This method is based on vector representations of scanpaths (i.e., in x and y space) and preserves a number of aspects, including the position and duration of fixations, the shape of scanpaths, and the length and direction of scanpath saccades. A more detailed technical description of this method is given by Jarodzka et al. (2010) and Dewhurst et al. (2012). This method computes five similarity measures for the paired fixation and saccade vectors of two given scanpaths, and these measures are then averaged to give five similarity scores. These five MultiMatch measures compare the vector, length, direction, position, and duration of two scanpaths (Fig. 3). Since the similarity score is calculated differently for each of the five measures, absolute score values cannot be compared across measures. However, the distributions of these similarity scores within the same measure for the control and experimental groups indicate whether one group had greater scanpath variability than the

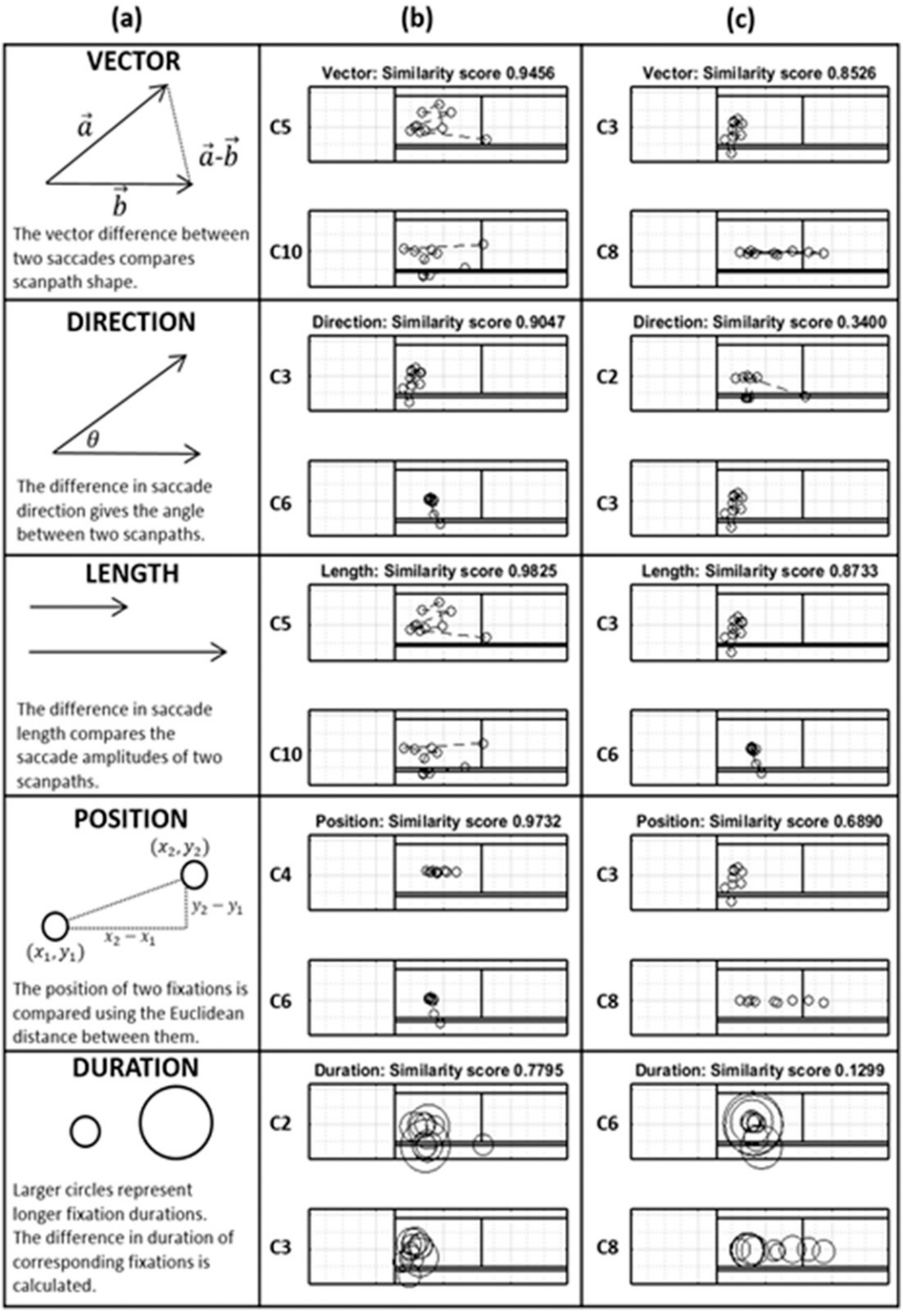

FIG. 3. The (a) five MultiMatch measures with corresponding examples of scanpaths that have relatively (b) higher and (c) lower similarity scores for two control group participants. [Adapted from Dewhurst et al. (2012).]

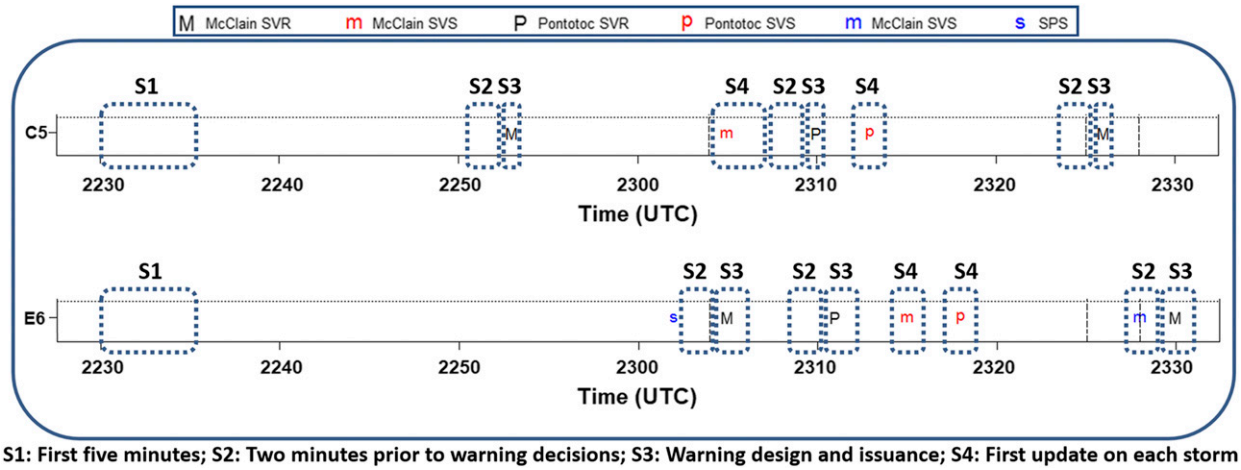

FIG. 4. Examples of the four defined stages occurring within control group participant C5's and experimental group participant E6's warning decision process. With the exception of the first 5 min, the general duration of each stage lasted 1–3 min. Severe thunderstorm (SVR) issuance times are shown for the McClain storm (M in black) and Pontotoc storm (P in black), and first SVS issuance times are shown for the McClain storm (m in red) and Pontotoc storm (p in red). Additional issuance decisions potentially impacting analyzed stages include the issuance of a special weather statement (s in blue) and the issuance of a second SVS for the McClain storm (m in blue). Vertical dashed lines indicate the timing of severe weather reports (see Fig. 5).

other. If greater scanpath variability in one group was observed, forecasters' video and retrospective recall data helped to explain why these differences occurred within the context of their warning decision processes.

4) DEFINED STAGES

Previous PARISE studies have observed that when working weather events in simulated real time, there are clear stages in the warning decision process that are common among all forecasters. To better understand the differences in forecasters' cognitive processes during times in which they are engaged in the same task, we chose to focus our analysis of the eye movement data during four stages: 1) the first 5 min of the case, 2) the 2 min prior to warning decisions, 3) the warning issuance process, and 4) the first update on the McClain and Pontotoc storms. The timing of these stages for all participants was identified using video and retrospective recall data (i.e., when forecasters opened WarnGen and recalled reaching a decision point) (Fig. 4), and their corresponding eye movement data were extracted for analysis. All forecasters issued a severe thunderstorm warning at least once during the McClain storm and once during the Pontotoc storm (Fig. 5). Eleven control and 10 experimental group participants also issued a second severe thunderstorm warning on the McClain storm (Fig. 5). Given that these were major warning decisions across both groups, the warning issuance process for each of these three decisions is included in the analysis. The timing of participants' warning decisions determined what storm processes could be viewed in the 2 min prior to these decision points. Finally, updates to these warnings were completed through the issuance of severe weather statements (SVSs). Some forecasters issued many more SVSs than others, but 11 participants in the experimental group and all participants in the control group issued at least one SVS on the McClain storm, and six participants in each group issued at least one SVS on the Pontotoc storm (Fig. 5). For the fourth stage, we therefore focused on the first SVS issuance for each of these storms.

For each of these stages, the participants' fixation count and mean fixation duration were calculated, and the five MultiMatch measures were computed for all possible participant scanpath combinations within each group. Of particular interest were the differences identified between the control and experimental groups' fixation and MultiMatch measures. Once these differences were identified, forecasters' video and retrospective recall data were used to provide context and explanation for why they occurred. These qualitative data were useful for determining whether differences in eye movements were due to forecasters' use of different radar update speeds or due to other factors, and they enabled an analysis of whether differences occurred at a group level or whether they were due to specific forecasters within the group.

3. Results

a. *First 5 min*

The first 5 min characterizes a time in which forecasters were busy loading their radar data and familiarizing

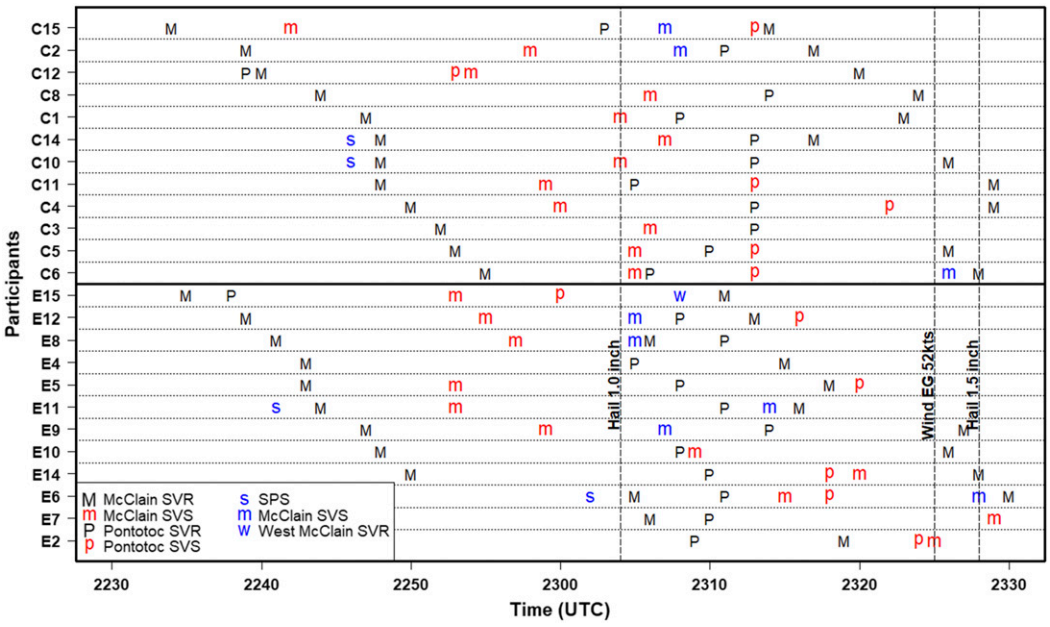

FIG. 5. (top) Control and (bottom) experimental group participants' warning products issued during the weather scenario. Markers are the same as in Fig. 4. Additionally, the SVR issuance time for the development west of the McClain storm is shown (w in blue).Vertical dashed lines indicate the timing of the severe weather reports associated with the McClain storm.

themselves with the weather scenario. Video and retrospective recall data show that forecasters in both groups spent much of their time sampling the reflectivity profiles of the McClain and Pontotoc storms, frequently moving back and forth between the two storms while climbing in elevation for vertical comparison. The eye fixation measures of the control (5-min PAR updates) and experimental (1-min PAR updates) groups reflect this observed behavior. Attention was given primarily to the reflectivity AOI, with the mean fixation count in this AOI exceeding that of any other AOI [$\text{Count}_{\text{con}}$(std dev) = 401 (34) and $\text{Count}_{\text{exp}}$(std dev) = 377 (77), where ''con'' is the control group, ''exp'' is the experimental group, and std dev stands for standard deviation]. The second highest mean fixation count for both groups occurred within the velocity AOI [$\text{Count}_{\text{con}}$(std dev) = 128 (69) and $\text{Count}_{\text{exp}}$(std dev) = 119 (53)]. Only the deeper McClain storm was visible at higher elevations, and for most forecasters a choice was made to prioritize attention on this storm. These observations led one participant in each group to issue a severe thunderstorm warning on the McClain storm because they thought that the storm was ''Going to take off'' after seeing ''65 dBZ above 25 kft'' (C15) and ''50 dBZ above 30 kft'' (E15). One additional participant in the experimental group (E5) also decided to be ''Proactive to be ready to issue a warning'' and therefore used this time to prepare a similar warning (Fig. 5). None of the other forecasters, however, visited the WarnGen AOI during this time.

Both groups' scanpaths were relatively more similar during these first 5 min compared to the later defined stages (Fig. 6). Differences in the groups' similarity scores for four of the five MultiMatch dimensions were not statistically significant, indicating a comparable level of variability in forecasters' scanpaths within each group. However, the groups did differ with respect to fixation duration (p value < 0.001), with the experimental group's lower similarity scores indicating more differences in their processing of these data (Fig. 6e). The experimental group's larger variation in mean fixation duration was most evident within the reflectivity and especially velocity [Dur_{con}(std dev) = 395 ms (48 ms) and Dur_{exp}(std dev) = 486 ms (132 ms)] AOIs, where the experimental group's spread in fixation duration was notably greater than that of the control group.

b. 2 min prior to warning decision

Forecasters' eye movements in the 2 min preceding a warning decision were analyzed for up to three occasions. No differences in fixation measures (not shown) or any of the five MultiMatch similarity scores were found to be statistically significant between the groups prior to the first warning on the McClain storm (Fig. 6). For most

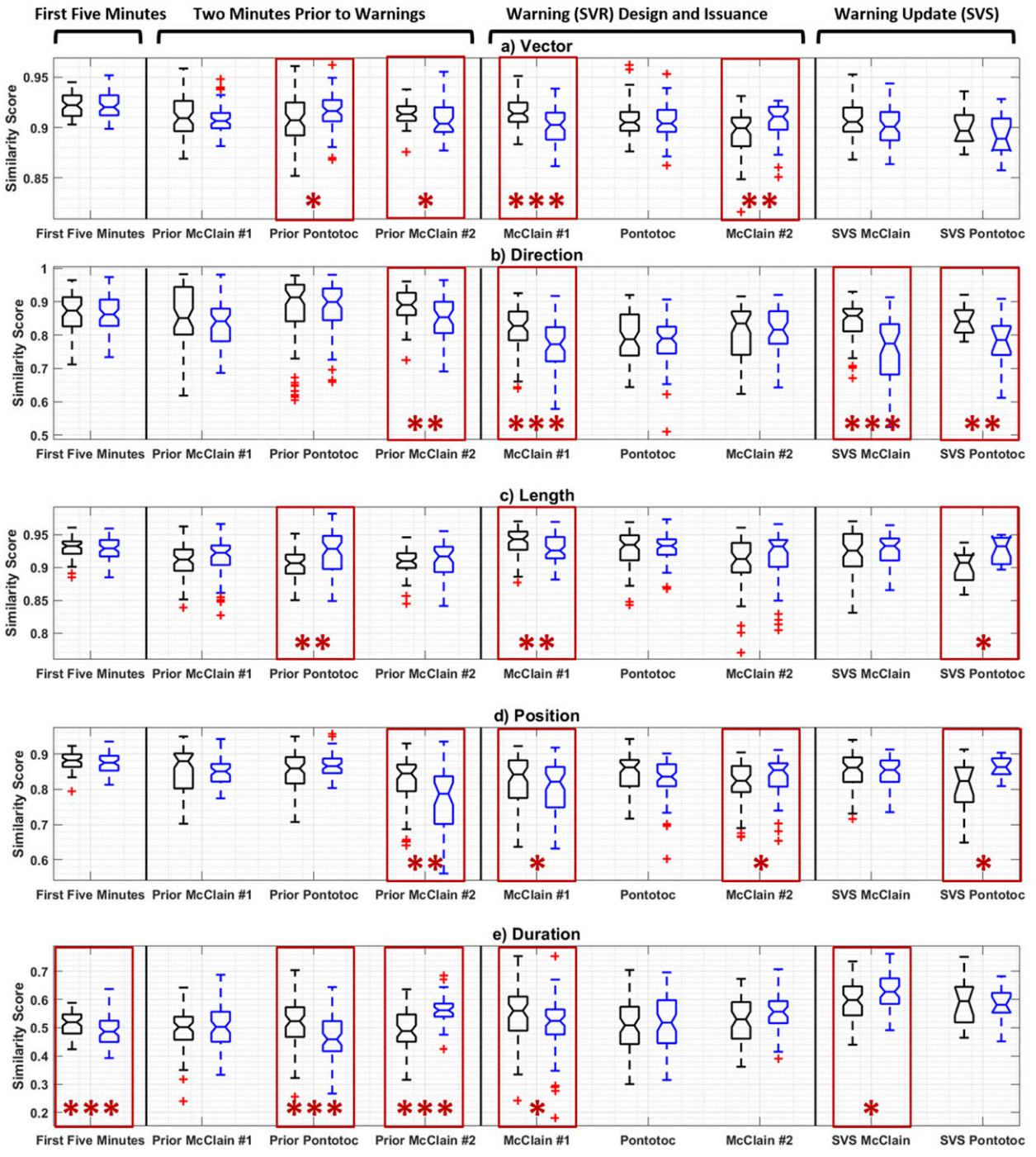

FIG. 6. Boxplot distributions of similarity scores for the five MultiMatch measures: (a) vector, (b) direction, (c) length, (d) position, and (e) duration for the control group (left position, black) and experimental group (right position, blue). Red boxes indicate distributions that are statistically significantly different according to the Wilcoxon–Mann–Whitney rank-sum test (* indicates a p value < 0.05, ** for a p value < 0.01, and *** for a p value < 0.001). Nonoverlapping notches also indicate strong evidence for differing medians. Red crosses (+) indicate outlier values that are less (greater) than 1.5 times the lower (upper) quartile.

participants, video data showed that interrogation continued in a manner similar to the first 5 min, such that fixations in the reflectivity AOI were 3–4 times as frequent as those in the velocity AOI. However, most forecasters expressed that they were "Concerned with the McClain cell" (e.g., C5) after seeing an increasing

trend in reflectivity values and a three-body scatter spike indicative of large hail associated with this storm. With these observations, forecasters shifted their attention to the McClain storm in the 2 min leading up to their decisions to warn, resulting in reduced references and attention given to the Pontotoc storm in the retrospective recall and video data.

Although fixation measures between the control and experimental groups were also not statistically significantly different in the 2 min prior to the Pontotoc warning (not shown), greater variability in the control group's scanpaths was observed in the vector and length MultiMatch dimensions as a result of two participants' anomalous scanpaths (Figs. 6a,c). Video data showed that the lower vector similarity scores were due to C12's chosen method for navigating through the radar data. While C12 preferred to click on icons located in the control AOIs (Fig. 7a), all other forecasters followed the taught method of toggling with computer keys. C15 was predominantly responsible for the lower length similarity scores because of their decision (unlike other forecasters) to focus interrogation only on the reflectivity AOI since their "Objective [was] to find hail cores aloft" (Fig. 7b). The overall shape of C12's scanpath and the shorter saccades belonging to C15's scanpath prior to the Pontotoc storm warning decision were visibly different than that of C10's. Participant C10 spent time "Looking for increases in velocity at 0.5°" and "Checking 70 dBZ" heights, and their scanpath (Fig. 7c) was a more typical representation of how forecasters interrogated the radar data prior to the Pontotoc warning decision. As this representative gaze plot shows, although forecasters' attention was distributed heavily within the reflectivity AOI prior to the Pontotoc storm warning, they also tended to check the velocity AOI. Both the video and retrospective recall data show that forecasters attended to the velocity AOI to analyze not only low-level wind signatures like C10, but also midlevel rotation and storm divergence (Fig. 7c).

While just two participants' unusual fixation patterns explained the control group's lower scanpath similarity prior to the Pontotoc warning, more prominent group differences occurred prior to the second McClain warning decision. On average, participants in the experimental group fixated twice as often in the WarnGen AOI than those in the control group [$\text{Count}_{\text{con}}$(std dev) = 27 (37) and $\text{Count}_{\text{exp}}$(std dev) = 62 (81)], while participants in the control group fixated more frequently within the velocity AOI [$\text{Count}_{\text{con}}$(std dev) = 48 (31) and $\text{Count}_{\text{exp}}$(std dev) = 33 (29)] but for a statistically significant shorter mean duration [Dur_{con}(std dev) = 372 ms (81 ms) and Dur_{exp}(std dev) = 452 ms (49 ms)] (p value = 0.0133). The higher velocity AOI fixation count corresponds to the control group participants' more frequent observations of the McClain storm's strengthening low-level wind signatures, such that participants were seeing "Downburst winds starting to spread out from the core" (C11) and "Higher winds along the northern flank… one pixel 34 kts" (C5) (1 kt = 0.51 m s^{-1}). Experimental group participants' greater use of WarnGen largely explains their statistically significant lower similarity scores for four of the five MultiMatch dimensions (Fig. 6). For example, E11's low similarity scores were due to spending much of these 2 min issuing a cancellation on the first McClain warning, having previously seen a downward trend in the reflectivity core (Fig. 8a). Following this cancellation, he "Noticed a gigantic three body scatter spike coming off that core that had 50 dBZ at 32 kft," and quickly decided to issue a second warning on this storm. In addition to E11's cancellation, observations of increasing reflectivity values aloft (and an associated updraft pulse) coupled with a storm report prompted E6 (Fig. 8b) and E8 to update the first McClain storm warning during these 2 min. E15's use of WarnGen during this time was a result of his decision to issue a warning on storm development to the west of the McClain storm given the strengthening 1-min trends in its reflectivity core (Fig. 5). Unlike these four participants in the experimental group, others within this group used their time to focus only on the radar data and produced a scanpath evidently different from those that carried out WarnGen-based tasks (e.g., E14; see Fig. 8c). Although these other experimental group participants checked the Pontotoc storm intermittently, most of their time was spent on the McClain storm "because it had various reports and the warning [was] coming close to expiration" (E14).

c. *Warning issuance process*

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

Although forecasters' fixation measures were comparable across both groups during the issuance of the first McClain warning, four participants' deviation from the typical issuance routine resulted in statistically

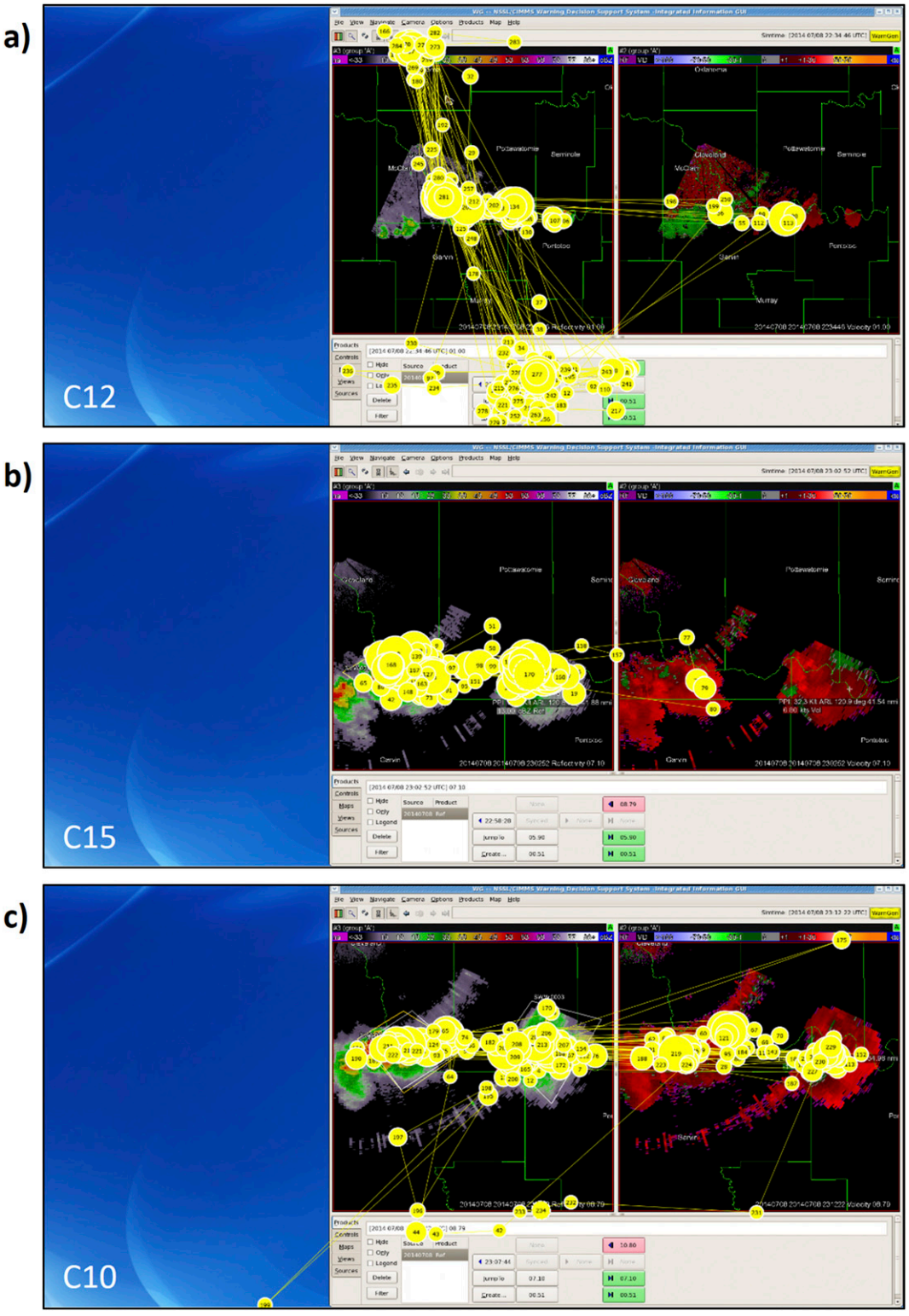

FIG. 7. Gaze plots depicting the scanpaths of participants (a) C12, (b) C15, and (c) C10 in the 2 min prior to the Pontotoc storm warning decision. Circles represent fixations, the circle center identifies the fixation location, and the circle size characterizes the fixation duration. Lines between fixations represent the corresponding saccades. The background screenshot is the final frame from the period depicted.

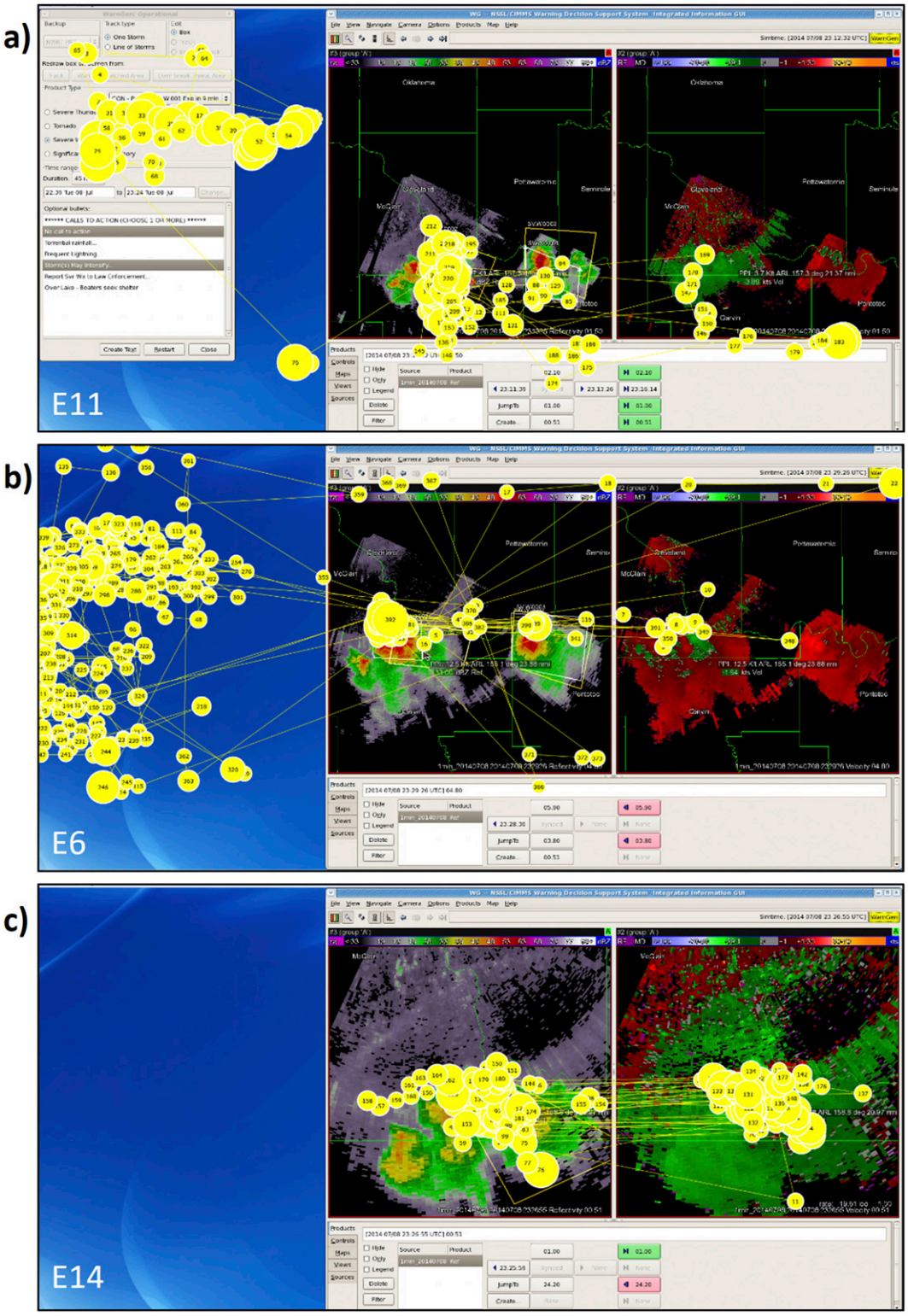

FIG. 8. Gaze plots depicting the scanpaths of participants (a) E11, (b) E6, and (c) E14 in the 2 min prior to the second McClain storm warning decision. Note that the WarnGen tool could be toggled on and off at any time and sometimes did not appear in the final screen capture.

significant lower scanpath similarity scores within the experimental group for all five MultiMatch dimensions (Fig. 6). For example, despite making a decision to warn, E2 reported that he did not feel the urgency to issue the warning given that the ''Situation was not rapidly evolving'' and he could ''Afford to spend time on [the] warning product [to get a] good handle on what's going on'' (Fig. 9a). While issuing the first McClain storm warning, E2 spent considerably more time than other forecasters watching storm trends, ensuring that the Pontotoc storm did not require his attention, and as he reported, ''Nitpicking small details.'' Similarly, while designing the warning, E12 ''Spent time thinking about decisions'' by analyzing trends in radar data and carefully considering what threats to include in the warning, which call to actions to select, and for how long the warning should be issued (Fig. 9b). Additionally, two forecasters struggled with technical disruptions when issuing the warning. E11 struggled to set the polygon correctly because he ''[Couldn't] fine tune counties as much as [he would] like,'' while E14 found that the polygon ''Kept snapping around on [him],'' causing him to switch between the reflectivity and WarnGen AOIs frequently and to have more broadly distributed fixations across the reflectivity AOI after repeatedly readjusting the vertices (Fig. 9c).

The few technical challenges observed during the issuance of the first McClain warning did not arise during the Pontotoc warning issuance and thus did not reduce the scanpath similarity among participants. Furthermore, the majority of participants' decisions to issue this warning were prompted within 5–10 min of receiving the first hail report and observing increased reflectivity values as the ''[Pontotoc] storm continues to get bigger'' (E10). The timing and reasoning of the Pontotoc storm warning was therefore much more similar than for the first McClain storm warning (Fig. 5). It is then unsurprising that forecasters followed the routine warning issuance process for the Pontotoc storm, and no statistically significant differences between the control and experimental groups' fixation measures or MultiMatch dimensions were observed (Fig. 6).

For most participants, the final warning was issued again on the McClain storm (Fig. 5). Unlike the first McClain warning, participants' scanpaths in the experimental group during this second issuance were more similar to one another than participants' scanpaths in the control group (Fig. 6). The scanpaths of three participants in the control group explain why the vector and position similarity scores were statistically significantly lower for this group. First, despite most other participants thinking that the McClain storm continued to pose a severe weather threat, C4 was ''Not impressed with the storm'' and ''Reluctantly'' decided to issue the second McClain storm warning after receiving all storm reports (Fig. 5). He zoomed into the McClain storm during this issuance and transitioned between the reflectivity and WarnGen AOIs only once (Fig. 10a). This single transition is an important aspect of C4's scanpath because it was more typical for forecasters to transition between these two AOIs multiple times during warning issuance. Like C4, C2 also ''Did not think the storm was severe enough to warn on again.'' However, the first hail report associated with the McClain storm prompted C2 to hesitantly issue a second warning given that his first McClain storm warning was issued early in the case and would soon be expiring (Fig. 5). C2's hesitance was evident in his numerous revisits to the reflectivity AOI to sample the magnitude of the high-reflectivity core while creating the warning. This behavior resulted in many more transitions between the reflectivity and WarnGen AOIs than what was typical of other participants (Fig. 10b). The third control group participant that presented an unusual scanpath was C10. While the issuance of the second McClain warning was a quick process for this participant, he had previously noted ''Increasing values in velocity'' and, therefore, visited the velocity AOI to monitor these data while designing the warning (Fig. 10c). If at all, video data show that most other participants only glanced at the velocity AOI during this warning issuance.

d. Warning update process

The timing and reasoning of the first update to the McClain storm warning was more varied among participants in the experimental group than those in the control group. Whereas retrospective recall data showed that a storm report drove more than half of the control group participants' decision to issue this SVS, most experimental group participants issued this update because of ''maintenance reasons,'' such as altering the expected weather threat based on radar observations, trimming areas of the warning polygon, or simply providing a continuation of the warning. The experimental group's greater spread in fixation counts within the reflectivity [$\text{Count}_{\text{con}}$(std dev) = 42 (30) and $\text{Count}_{\text{exp}}$(std dev) = 40 (40)] and WarnGen [$\text{Count}_{\text{con}}$ (std dev) = 124 (59) and $\text{Count}_{\text{exp}}$(std dev) = 146 (96)] AOIs, along with their statistically significant lower direction similarity scores (Fig. 6), illustrates their more variable scanpaths during this warning update compared to the control group. As observed in the video data, the experimental group's lower direction similarity scores occurred as a result of participants who either updated the warning with an unusually quick or an unusually extended process. For example, while a couple of experimental group participants

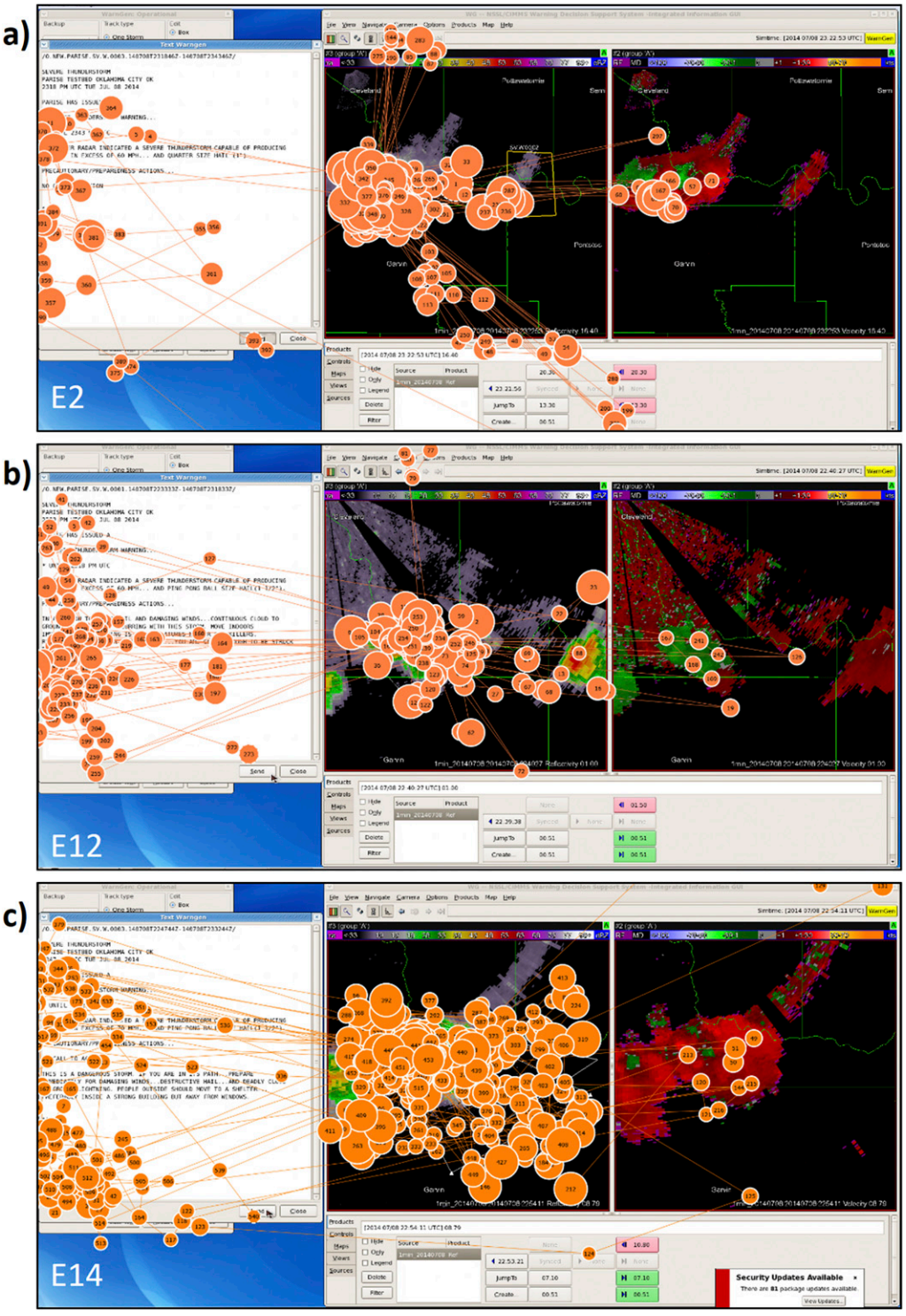

FIG. 9. Gaze plots depicting the scanpaths of participants (a) E2, (b) E12, and (c) E14 during the issuance of the first McClain storm warning.

issued the SVS without changing any aspect of the warning because they thought there had "Been little change in the storm overall" (e.g., E10; see Fig. 11a), others spent considerable time assessing the radar data, updating the expected weather threat, and carefully adjusting the polygon vertices (e.g., E5; see Fig. 11b). These

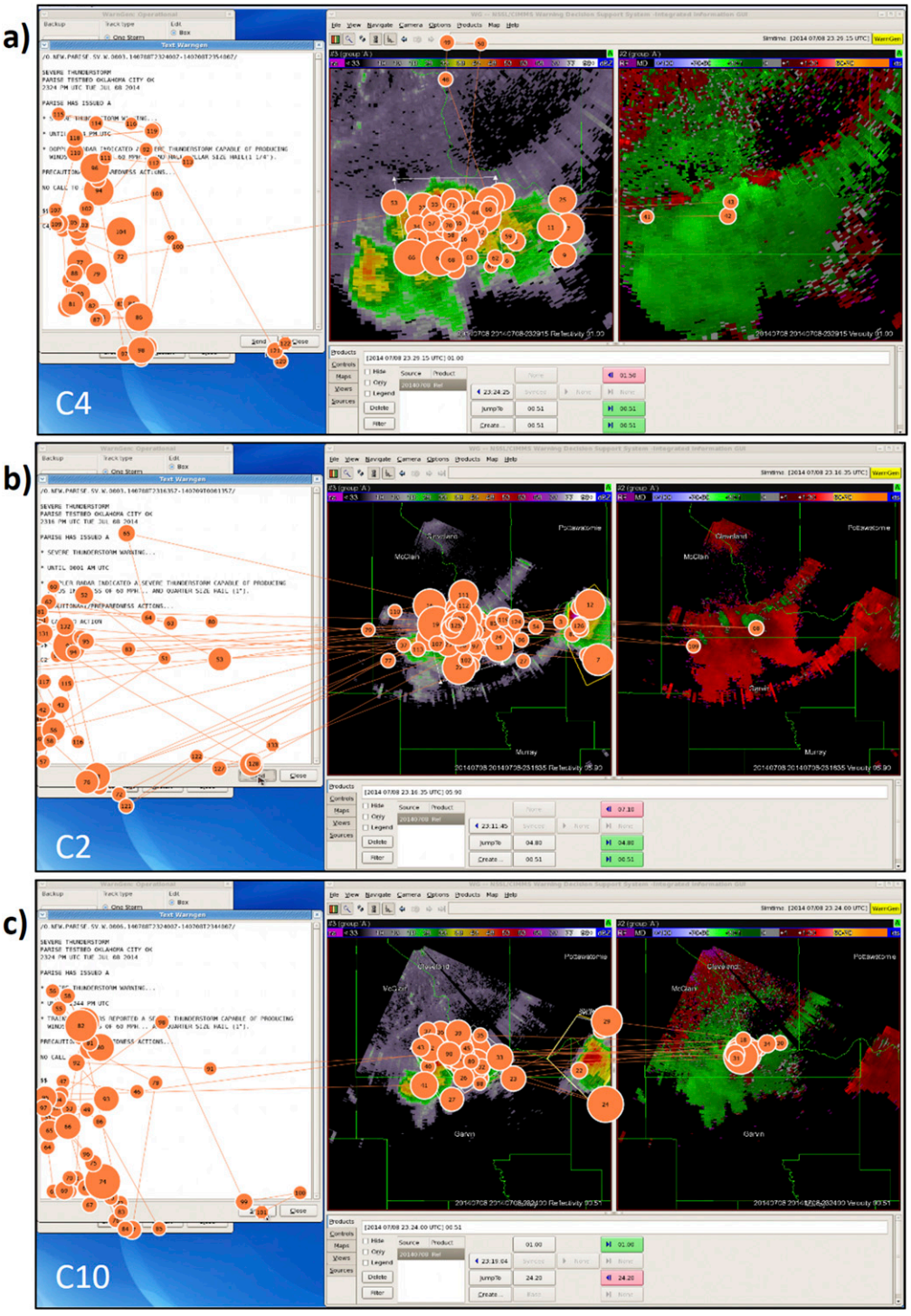

FIG. 10. Gaze plots depicting the scanpaths of participants (a) C4, (b) C2, and (c) C10 during the issuance of the second McClain storm warning.

contrasting warning update processes were not observed in the control group; rather, all participants in the control group changed at least one aspect of the warning.

Six participants in each group chose to issue an SVS on the Pontotoc storm (Fig. 5), and since no severe weather was reported for this storm, all updates were based only

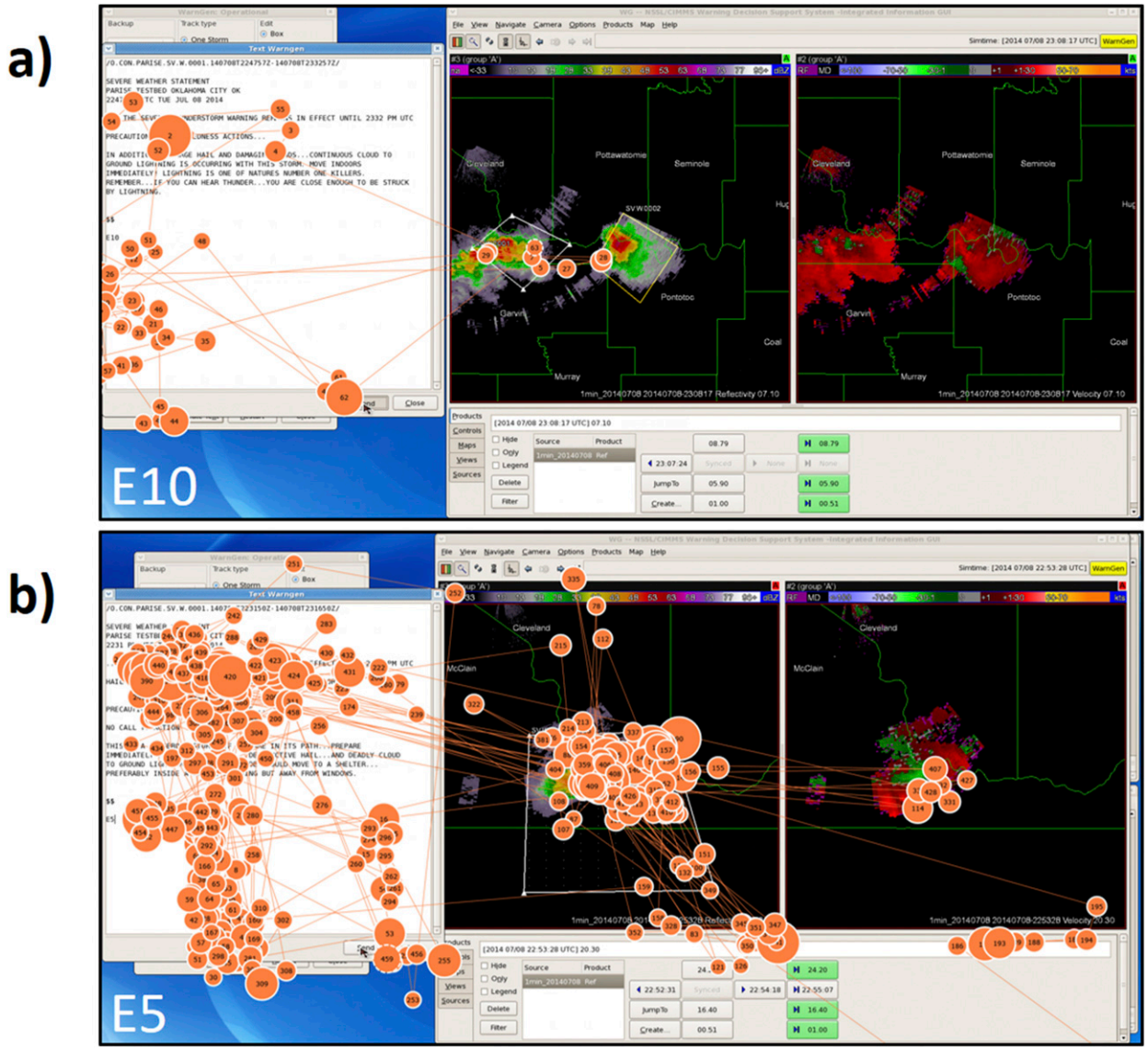

FIG. 11. Gaze plots depicting the scanpaths of participants (a) E10 and (b) E5 during the first McClain storm warning update.

on maintenance reasons. Video data showed that the experimental group's statistically significantly lower direction similarity scores were primarily a result of E15's more careful adjustment of the warning polygon vertices and lack of editing within the text portion of the WarnGen AOI compared to other participants in this group. While the control group was more similar with respect to scanpath direction, they were statistically significantly less similar than the experimental group in the length and position MultiMatch dimensions (Fig. 6). The lower length similarity scores were due to participants C4 (Fig. 12a) and C15 (Fig. 12b) focusing their attention predominantly in the reflectivity and WarnGen AOIs, respectively. C4 issued this update to trim the warning polygon, while C15 wanted to add text in the warning to communicate the expected hail threat, which resulted in C4 and C15 having the fewest fixations in the WarnGen and reflectivity AOIs out of the control group, respectively. Finally, the statistically significantly lower position scores in the control group were a result of C5's sporadically placed fixations, which according to the video data likely resulted from his eye gaze darting between the keyboard and computer screen while editing warning text.

e. Differences in duration

Unlike the vector, direction, length, and position MultiMatch measures, similarity in fixation duration is difficult to visualize in gaze plots, and thus it is

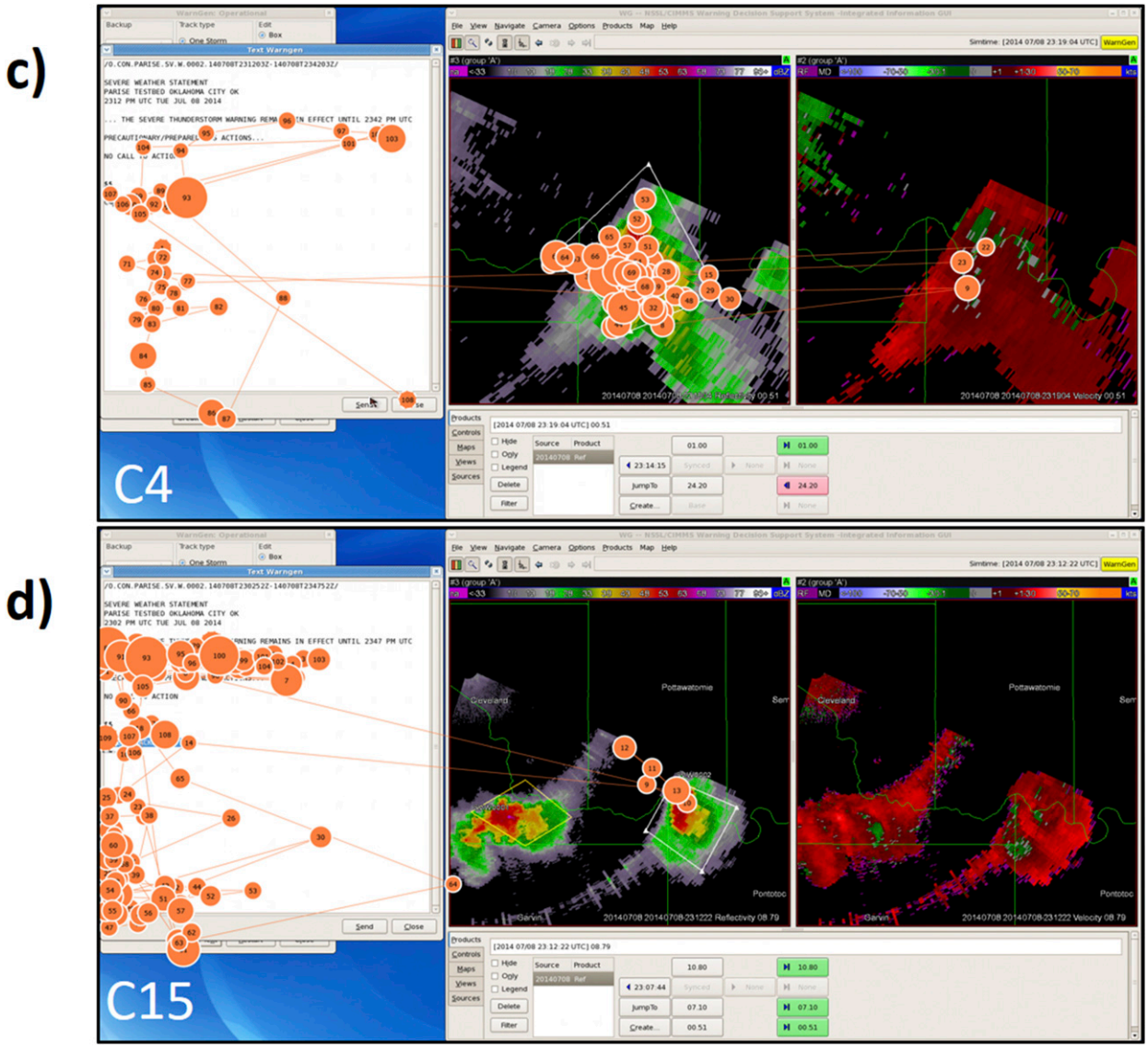

FIG. 12. Gaze plots depicting the scanpaths of participants (a) C4 and (b) C15 during the first Pontotoc storm warning update.

challenging to compare between forecasters. When focused on a piece of information, a person's fixation duration is indicative of their level of engagement and effort in extracting and processing it (Poole and Ball 2006; Bojko 2013). In each of the four defined stages, the difference in fixation duration similarity scores among control group participants and among experimental group participants was statistically significant at least once (Fig. 6e). However, in only one of these instances did the group with statistically significant *lower* duration similarity scores also have statistically significant *lower* similarity scores in other MultiMatch dimensions (Fig. 6). In the other instances, either no statistically significant difference was found for the vector, direction, length, or position dimensions, or the group that experienced statistically significantly *more* variation in duration was the one to experience statistically significantly *less* variation in other dimensions (Fig. 6). This result demonstrates that even when forecasters' placement of and transition between fixations is similar, how intently they focus on information can still vary.

4. Discussion and conclusions

This eye-tracking experiment provided an opportunity to demonstrate a mixed-methods approach to collecting forecasters' eye movement data, retrospective

recalls, and computer display activity videos in a simplified warning scenario. Working in experiment conditions, we found that forecasters were comfortable knowing that their eyes were being tracked, they responded well to sitting relatively still throughout the study, and were generally accepting of the need to constrain both the amount of data available for interrogation and the flexibility in data display. To allow for some real-life interrogation practices, forecasters were given the ability to navigate through radar data in time and space and reposition storms within AOIs through zooming and panning. The most notable limitation in providing these viewing abilities was that all forecasters toggled frequently in time and height, so analysis of eye movement data alone could only inform on the type of information that was being viewed instead of what was specifically being viewed by a forecaster at any given time. Furthermore, while these viewing abilities did not affect forecasters' bulk fixation measures, it is possible that the MultiMatch dimensions were impacted. In light of the limitations resulting from these viewing abilities, interpretation of the eye movement results alongside the video and retrospective recall data was essential for determining what portion of the storm forecasters were viewing (e.g., low-level velocity vs storm-top divergence or magnitude of the reflectivity core at the lowest elevation vs at 20 kft) and what their related thoughts were. Providing meaningful explanation to all quantified differences in forecasters' eye movements required joint consideration of all of the collected data throughout the analysis process.

A more specific goal of this eye-tracking experiment was to use forecasters' eye movement data alongside the qualitative data to identify and explain differences in their warning decision processes as a result of using different radar update speeds. However, we found that the fixation measures were generally comparable throughout the four defined stages due to forecasters' similar distributions of attention regardless of whether 1- or 5-min PAR updates were used. In retrospect, we believe that the chosen weather scenario strongly influenced this result. Most forecasters had stated prior to beginning the case that they expected the weather threat to be primarily hail and secondarily wind. It is then unsurprising that forecasters focused predominantly in the reflectivity AOI, switching often between the persistent McClain and Pontotoc storms, with more intermittent checking in the velocity AOI. The threat expectation for these slow-moving multicell storms did not change throughout the case, and thus this interrogation pattern was maintained for much of the hour. It remains to be seen whether differences in the fixation measures of forecasters using 1- and 5-min PAR data would be greater if a more dynamic event with changing weather threats was presented [e.g., such as the case presented in Wilson et al. (2016)].

Forecasters' eye movements were further analyzed using the MultiMatch scanpath comparison algorithm (Jarodzka et al. 2010). Although other scanpath comparison methods also maintain the temporal ordering of fixation sequences (e.g., Levenshtein 1966; Cristino et al. 2010; Jarodzka et al. 2010; Anderson et al. 2015), the MultiMatch method was chosen because it compares scanpaths at a finer spatial scale. With this higher spatial resolution, we found that the MultiMatch method detected statistically significant differences in the variability of forecasters' scanpaths in the control and experimental groups. The video and retrospective recall data revealed that these differences occurred as a result of the sensitivity of the MultiMatch method to how individual forecasters interacted with the user interface, tackled technological glitches in WarnGen, and approached tasks differently.

The findings from this experiment suggest that applications of eye-tracking and qualitative research methods together could be useful for other avenues of operational meteorology research. With new types of data and products being introduced to forecasters often, there are opportunities to use these methods to evaluate the learnability and usability of this new information (e.g., Jacob and Karn 2003). Given the highly dynamic display systems that forecasters use during operations, focused evaluations of simplified interfaces would first be necessary to analyze and interpret basic eye movement. This research would support the development of user-friendly interfaces that display information in an effective manner and improve human–computer interactions for operational meteorologists. We are hopeful that future applications of these research methods, such as the above-suggested usability studies, will expand our understanding of forecasters' cognition and act to support their important role within the weather enterprise.

Acknowledgments. We thank the 30 NWS forecasters and their respective Weather Forecast Offices for their participation in the 2015 PARISE. Additionally, we appreciate Darrel Kingfield for providing the temporally degraded PAR data, Jeff Brogden for developing the WarnGen tool for WDSS-II, Chris Carter and Robert Coggins for their ongoing technical support during the eye-tracking experiment, Greg Schoor for participating in the pilot study, and Gabe Garfield for preparing the 2015 PARISE prebriefing videos. Thank you to Jared Allen and David Schvartzman for providing a helpful review of this article and to three reviewers for their thoughtful feedback and suggestions during the review process. Funding was provided by the NOAA/Office of Oceanic

and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. The contents of this paper do not necessarily reflect the views or official position of any organization of the U.S. government.

REFERENCES

Al-Moteri, M. O., M. Symmons, V. Plummer, and S. Cooper, 2017: Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Comput. Hum. Behav.*, **66**, 52–66, https://doi.org/10.1016/j.chb.2016.09.022.

Anderson, N. C., F. Anderson, A. Kingstone, and W. F. Bischof, 2015: A comparison of scanpath comparison methods. *Behav. Res.*, **47**, 1377–1392, https://doi.org/10.3758/s13428-014-0550-3.

Bertram, R., J. Kaakinen, F. Bensch, L. Helle, E. Lantto, P. Niemi, and N. Lundbom, 2016: Eye movements of radiologists reflect expertise in CT study interpretation: A potential tool to measure resident development. *Radiology*, **281**, 805–815, https://doi.org/10.1148/radiol.2016151255.

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

Bowden, K. A., and P. L. Heinselman, 2016: A qualitative analysis of NWS forecasters' use of phased-array radar data during severe hail and wind events. *Wea. Forecasting*, **31**, 43–55, https://doi.org/10.1175/WAF-D-15-0089.1.

——, ——, D. M. Kingfield, and R. Thomas, 2015: Impacts of phased-array radar data on forecaster performance during severe hail and wind events. *Wea. Forecasting*, **30**, 389–404, https://doi.org/10.1175/WAF-D-14-00101.1.

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

Calhoun, K. M., T. M. Smith, D. M. Kingfield, J. Gao, and D. J. Stensrud, 2014: Forecasters use and evaluation of real-time 3DVAR analyses during severe thunderstorm and tornado warning operations in the Hazardous Weather Testbed. *Wea. Forecasting*, **29**, 601–613, https://doi.org/10.1175/WAF-D-13-00107.1.

Clement, J., T. Kirstensen, and K. Grønhaug, 2013: Understanding consumers' in-store visual perception: The influence of package design features on visual attention. *J. Retailing Consum. Serv.*, **20**, 234–239, https://doi.org/10.1016/j.jretconser.2013.01.003.

Cristino, F., S. Mathôt, J. Theeuwes, and I. D. Gilchrist, 2010: ScanMatch: A novel method for comparing fixation sequences. *Behav. Res. Methods*, **42**, 692–700, https://doi.org/10.3758/BRM.42.3.692.

Dewhurst, R., M. Nyström, J. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist, 2012: It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behav. Res. Methods*, **44**, 1079–1100, https://doi.org/10.3758/s13428-012-0212-2.

Djamasbi, S., M. Siegel, and T. Tullis, 2010: Generation Y, web design, and eye tracking. *Int. J. Hum. Comput. Stud.*, **68**, 307–323, https://doi.org/10.1016/j.ijhcs.2009.12.006.

Drost, R., J. Trobec, C. Steffke, and J. Libarkin, 2015: Eye tracking: Evaluating the impact of gesturing during televised weather forecasts. *Bull. Amer. Meteor. Soc.*, **96**, 387–392, https://doi.org/10.1175/BAMS-D-13-00217.1.

Duchowski, A. T., 2002: A breadth-first survey of eye-tracking applications. *Behav. Res. Methods Instrum. Comput.*, **34**, 455–470, https://doi.org/10.3758/BF03195475.

Ericsson, K. A., and H. A. Simon, 1993: *Protocol Analysis: Verbal Reports as Data*. Revised ed. The MIT Press, 496 pp.

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

Giovinco, N. A., S. M. Sutton, J. D. Miller, T. M. Rankin, G. W. Gonzalez, B. Najafi, and D. Armstrong, 2015: A passing glance? Differences in eye tracking and gaze patterns between trainees and experts reading plain film bunion radiographs. *J. Foot Ankle Surg.*, **54**, 382–391, https://doi.org/10.1053/j.jfas.2014.08.013.

Goldberg, H. J., and A. M. Wichansky, 2003: Eye tracking in usability evaluation: A practitioner's guide. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, J. Hyönä, R. Radach, and H. Deubel, Eds., Elsevier, 493–516.

Goodman, S. J., and Coauthors, 2012: The GOES-R Proving Ground: Accelerating user readiness for the next-generation geostationary environmental satellite system. *Bull. Amer. Meteor. Soc.*, **93**, 1029–1040, https://doi.org/10.1175/BAMS-D-11-00175.1.

Hauland, G., 2008: Measuring individual and team situation awareness during planning tasks in training of en route air traffic control. *Int. J. Aviat. Psychol.*, **18**, 290–304, https://doi.org/10.1080/10508410802168333.

Heinselman, P. L., D. S. LaDue, and H. Lazrus, 2012: Exploring impacts of rapid-scan radar data on NWS decisions. *Wea. Forecasting*, **27**, 1031–1044, https://doi.org/10.1175/WAF-D-11-00145.1.

——, ——, D. M. Kingfield, and R. Hoffman, 2015: Tornado warning decisions using phased-array radar data. *Wea. Forecasting*, **30**, 57–78, https://doi.org/10.1175/WAF-D-14-00042.1.

Henderson, J. M., and F. Ferreira, 2004: *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Psychology Press, 414 pp.

Hervet, G., K. Guérard, S. Tremblay, and M. S. Chtourou, 2011: Is banner blindness genuine? Eye tracking internet text advertising. *Appl. Cogn. Psychol.*, **25**, 708–716, https://doi.org/10.1002/acp.1742.

Hoffman, R. R., 2005: Protocols for cognitive task analysis. Florida Institute for Human and Machine Cognition Rep., 108 pp. [DTIC ADA475456.]

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

Huey, E. B., 1908: *The Psychology and Pedagogy of Reading*. Macmillan, 469 pp.

Hvelplund, K. T., 2014: Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. *MonTI Special Issue—Minding Translation*, R. M. Martin, Ed., Publicaciones de la Universidad de Alicante, 201–224, http://www.e-revistes.uji.es/index.php/monti/article/view/1706/1489.

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

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

Just, M. A., and P. A. Carpenter, 1976: Eye fixations and cognitive processes. *Cognit. Psychol.*, **8**, 441–480, https://doi.org/10.1016/0010-0285(76)90015-3.

——, and ——, 1980: A theory of reading: From eye fixations to comprehension. *Psychol. Rev.*, **87**, 329–354, https://doi.org/10.1037/0033-295X.87.4.329.

Kang, Z., and S. J. Landry, 2014: Using scanpaths as a learning method for a conflict detection task of multiple target tracking. *Hum. Factors*, **56**, 1150–1162, https://doi.org/10.1177/0018720814523066.

——, and ——, 2015: An eye movement analysis algorithm for a multi-element target tracking task: Maximum transition-based agglomerative hierarchical clustering. *IEEE Trans. Hum. Mach. Syst.*, **45**, 13–24, https://doi.org/10.1109/THMS.2014.2363121.

Karstens, C., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, https://doi.org/10.1175/WAF-D-14-00163.1.

Komogortsev, O. V., D. V. Gobert, S. Jayarathna, D. H. Koh, and S. M. Gowda, 2010: Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Trans. Biomed. Eng.*, **57**, 2635–2645, https://doi.org/10.1109/TBME.2010.2057429.

Lakshmanan, V., T. Smith, G. J. Stumpf, and K. Hondl, 2007: The Warning Decision Support System–Integrated Information. *Wea. Forecasting*, **22**, 596–612, https://doi.org/10.1175/WAF1009.1.

Levenshtein, V., 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, **10**, 707–710.

Liversedge, S. P., and J. M. Findlay, 2000: Saccadic eye movements and cognition. *Trends Cogn. Sci.*, **4**, 6–14, https://doi.org/10.1016/S1364-6613(99)01418-7.

Manning, D. J., S. C. Ethell, and T. Donovan, 2004: Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *Br. J. Radiol.*, **77**, 231–235, https://doi.org/10.1259/bjr/28883951.

Noton, D., and L. Stark, 1971: Scanpaths in saccadic eye movement while viewing and recognizing patterns. *Vision Res.*, **11**, 929–942, https://doi.org/10.1016/0042-6989(71)90213-6.

Olsen, A., 2012: The Tobii I-VT fixation filter: Algorithm description. Tobii Technology, 21 pp., https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii-pro-i-vt-fixation-filter.pdf.

Poole, A., and L. J. Ball, 2006: Eye tracking in HCI and usability research: Current status and future prospects. *Encyclopedia of Human Computer Interaction*, C. Ghaoui, Ed., Idea Group, 211–219.

Rayner, K., 1998: Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.*, **124**, 372–422, https://doi.org/10.1037/0033-2909.124.3.372.

Romano Bergstrom, J. C., E. R. Olmsted-Hawala, and M. E. Jans, 2013: Age-related differences in eye tracking and usability performance: Website usability for older adults. *Int. J. Hum. Comput. Interact.*, **29**, 541–548, https://doi.org/10.1080/10447318.2012.728493.

Sherman-Morris, K., K. B. Antonelli, and C. C. Williams, 2015: Measuring the effectiveness of the graphical communication of hurricane storm surge threat. *Wea. Climate Soc.*, **7**, 69–82, https://doi.org/10.1175/WCAS-D-13-00073.1.

Smith, T. M., and Coauthors, 2014: Examination of a real-time 3DVAR analysis system in the Hazardous Weather Testbed. *Wea. Forecasting*, **29**, 63–77, https://doi.org/10.1175/WAF-D-13-00044.1.

——, and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Sullivan, J., J. H. Yang, M. Day, and Q. Kennedy, 2011: Training simulation for helicopter navigation by characterizing visual scan patterns. *Aviat. Space Environ. Med.*, **82**, 871–878, https://doi.org/10.3357/ASEM.2947.2011.

Tobii Technology, 2014: Tobii TX300 Eye Tracker. Tobii Technology User Manual, 44 pp., https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-tx300-eye-tracker-user-manual.pdf/?v=2.0.

Trafton, J. G., S. Marshall, F. Mintz, and S. B. Trickett, 2002: Extracting explicit and implicit information from complex visualizations. *Diagramatic Representation and Inference*, M. Hegarty, B. Meyer, and N. H. Narayanan, Eds., Lecture Notes in Computer Science, Vol. 2317, Springer, 206–220, https://doi.org/10.1007/3-540-46037-3_22.

Van de Merwe, K., H. Van Dijk, and R. Zon, 2012: Eye movements as an indicator of situation awareness in a flight simulator experiment. *Int. J. Aviat. Psychol.*, **22**, 78–95, https://doi.org/10.1080/10508414.2012.635129.

Wang, Q., S. Yang, M. Liu, Z. Cao, and Q. Ma, 2014: An eye-tracking study of website complexity from cognitive load perspective. *Decis. Support Syst.*, **62**, 1–10, https://doi.org/10.1016/j.dss.2014.02.007.

Wilson, K. A., P. L. Heinselman, and Z. Kang, 2016: Exploring applications of eye tracking in operational meteorology research. *Bull. Amer. Meteor. Soc.*, **97**, 2019–2025, https://doi.org/10.1175/BAMS-D-15-00148.1.

——, ——, and C. M. Kuster, 2017a: Considerations for phased-array radar data use within the National Weather Service. *Wea. Forecasting*, **32**, 1959–1965, https://doi.org/10.1175/WAF-D-17-0084.1.

——, ——, ——, and D. M. Kingfield, 2017b: Forecaster performance and workload: Does radar update time matter? *Wea. Forecasting*, **32**, 253–274, https://doi.org/10.1175/WAF-D-16-0157.1.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303, https://doi.org/10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.

Wood, G., K. M. Knapp, B. Rock, C. Cousens, C. Roobottom, and M. R. Wilson, 2013: Visual expertise in detecting and diagnosing skeletal fractures. *Skeletal Radiol.*, **42**, 165–172, https://doi.org/10.1007/s00256-012-1503-5.

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

Yu, C. S., E. E. M. Wang, W. C. Li, G. Braithwaite, and N. Greaves, 2016: Pilots' visual scan patterns and attention distribution during the pursuit of a dynamic target. *Aerosp. Med. Hum. Perform.*, **87**, 40–47, https://doi.org/10.3357/AMHP.4209.2016.