DOT/FAA/TC-14/16

Federal Aviation Administration
William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

Now You See Me, Now You Don't: Change Blindness in Pilot Perception of Weather Symbology

Ulf Ahlstrom, FAA Human Factors Branch, ANG-E25
Joel Suss, Spectrum Software Technology, Inc.

June 2014

Technical Report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. This document does not constitute Federal Aviation Administration (FAA) certification policy. Consult your local FAA aircraft certification office as to its use.

This report is available at the FAA William J. Hughes Technical Center's full-text Technical Reports Web site: http://actlibrary.tc.faa.gov in Adobe® Acrobat® portable document format (PDF).

1. Report No. DOT/FAA/TC-14/16	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Now You See Me, Now You Don't: Change Blindness in Pilot Perception of Weather Symbology		**5. Report Date** June 2014	
		6. Performing Organization Code ANG-E25	
7. Author(s) Ulf Ahlstrom, FAA Human Factors Branch Joel Suss, Spectrum Software Technology, Inc.		**8. Performing Organization Report No.** DOT/FAA/TC-14/16	
9. Performing Organization Name and Address Federal Aviation Administration Human Factors Branch William J. Hughes Technical Center Atlantic City International Airport, NJ 08405		**10. Work Unit No. (TRAIS)**	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Federal Aviation Administration Weather Technology in the Cockpit (WTIC) 800 Independence Avenue, S.W. Washington, DC 20591		**13. Type of Report and Period Covered** Technical Report	
		14. Sponsoring Agency Code ANG-C61	

15. Supplementary Notes

16. Abstract

Objective: The overarching goal of this study is to perform a human factors assessment of the effects of variations in cockpit weather symbology on General Aviation (GA) pilot symbol perception. **Background**: To support the Next Generation Air Transportation System (NextGen) program, ongoing efforts focus on the implementation and use of weather technologies and weather presentations. **Method**: Sixty instrument-rated GA pilots volunteered to participate in the study. We manipulated the independent variable Weather Presentations by presenting weather information under three different symbology modes. In Experiment 1, we assess pilot perception of METAR symbols during flight and assess how this affects flight behavior, cognitive engagement, and decision-making. In Experiment 2, we focus on pilot perception of time-stamps and weather symbols in a "change-detection" experiment. **Results**: The result shows that pilots (using different weather presentations) vary considerably in their overall perception of METAR symbol change during flight. The overall group detection ranges from a virtual blindness (25% detections) to a modest detection performance (62% detections). The result from the change-detection experiment shows that the detection accuracy varies greatly between different weather symbols and between different weather presentations. Although the average change-detection performance is high across all weather presentations for precipitation areas (on average, 89% to 94% correct detections), SIGMET areas (83% to 93%), and METAR symbols (83% to 91%), pilots are virtually blind to changes for lightning symbols (17% to 43%) and time-stamp information (13% to 20%). **Conclusion**: Weather presentation symbology affects pilots' perception of symbol change and cognitive engagement. Pilot performance varies credibly between different symbology renderings of the same weather data. **Applications**: This simulation is part of an ongoing assessment of the effects of weather-presentation symbology related to the optimization of weather presentations in cockpits.

17. Key Words Change Detection Cockpit Simulation Cognitive Engagement Visual Flight Rules Weather Symbology		18. Distribution Statement This document is available to the public through the National Technical Information Service, Alexandria, Virginia, 22312. A copy is retained for reference at the William J. Hughes Technical Center Library.	
19. Security Classification (of this report) Unclassified	**20. Security Classification (of this page)** Unclassified	**21. No. of Pages** 113	**22. Price**
Form DOT F 1700.7 (8-72)	**Reproduction of completed page authorized**		

THIS PAGE IS BLANK INTENTIONALLY.

Table of Contents

List of Illustrations

Tables **Page**

THIS PAGE IS BLANK INTENTIONALLY.

Acknowledgments

THIS PAGE IS BLANK INTENTIONALLY.

Executive Summary

Visual Flight Rules (VFR) flight into Instrument Meteorological Conditions (IMC) is a major safety hazard for General Aviation (GA) pilots. Because of this and other flight-related weather dangers, pilots need not only to assess weather information as part of their pre-flight planning but also to maintain good "weather situation awareness" while in flight. One increasingly popular method for receiving in-flight weather updates is the use of cockpit weather displays, including certified installed display systems and noncertified portable devices, such as tablets or cell phones. Using these devices, pilots can display a wide variety of weather information elements including precipitation, wind, lightning, echo top, aviation routine weather reports (METARs), and significant meteorological information (SIGMET).

Graphical weather presentations could, potentially, aid a pilot and contribute to enhance weather situation awareness during flight. However, previous research has found some negative effects on pilot behavior and decision-making from the use of graphical weather data. In some cases, the graphical weather information was displayed but not used, but in other cases the information was not used efficiently. One reason for the inefficient use of information could stem from the symbol color and shapes used to display the graphical weather information. For example, if a symbol color is not easily detectable against the display background, pilots may fail to perceive important weather information. This raises the question of how easy it is for pilots to see different symbol colors and shapes and how symbol variations affect pilots' awareness of the presence of symbols and symbol color changes during flight.

In this study, we evaluate whether variations in weather symbols (i.e., differences in symbol shapes and colors) affect pilots' ability to detect weather changes and pilots' flight planning and flying behavior. During the study, pilots performed two different tasks: Task 1 and Task 2. In Task 1, pilots flew a cockpit simulator that was configured to simulate a single-engine aircraft. For this task, we assess how variations in symbol colors and shapes affect pilots' ability to detect symbol changes for deteriorating weather conditions at airports. The initial weather conditions were VFR, allowing the pilots to see out the window while flying. At locations where the symbols changed, indicating deteriorating conditions, pilots were required to fly using instrument flight rules (IFR). In Task 2, we assess the extent to which differences in symbol shapes and colors affect how easily pilots can detect symbol changes in static images.

Cockpit weather presentations

We compare three weather symbol sets that varied in the shapes and colors used to represent weather information in the cockpit (see Figure 1). We chose these symbol sets because they are representative of the variety of symbologies used in currently available commercial weather-display products.
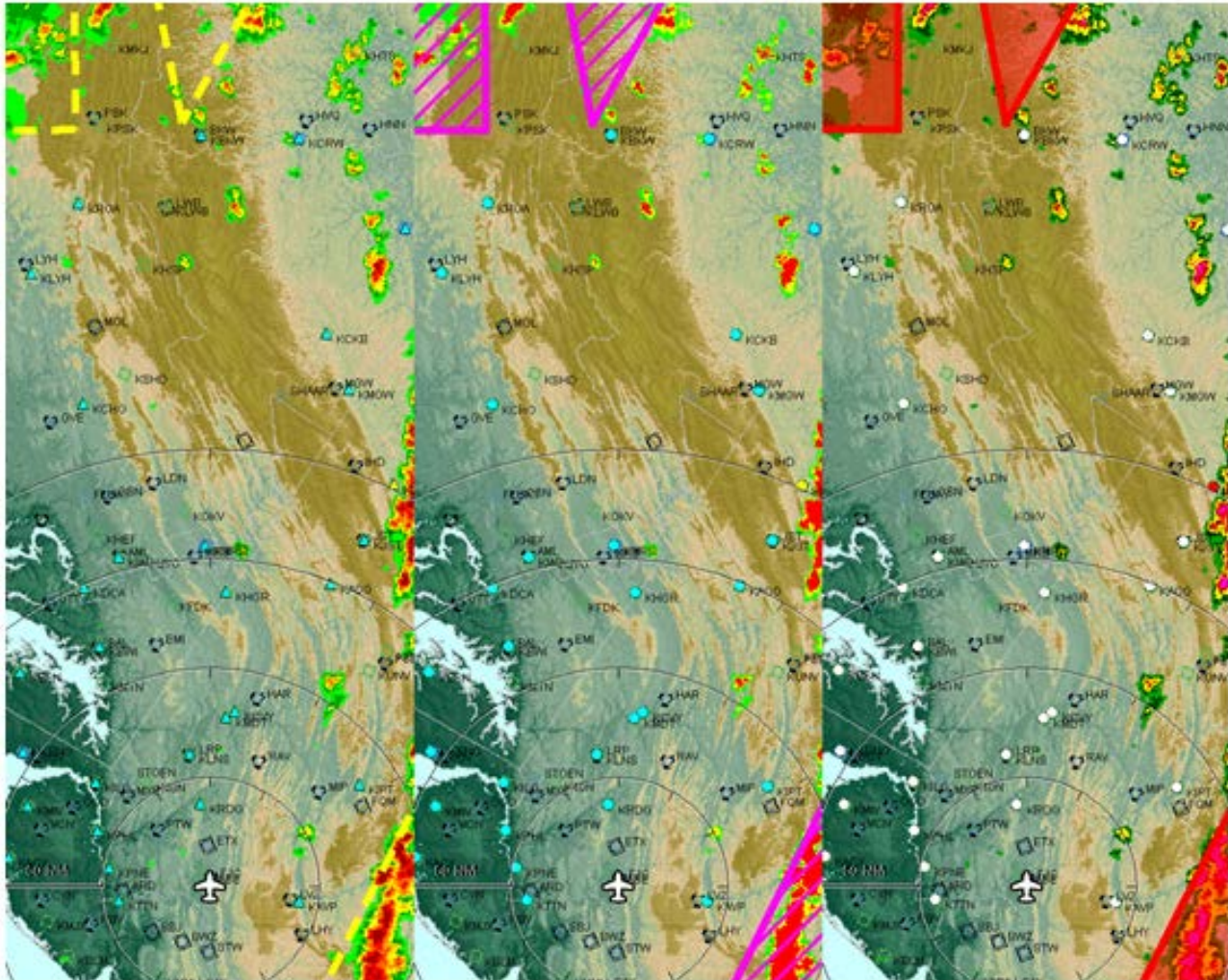
Figure 1. The three weather presentations used during the cockpit simulation flights. The METAR symbols are displayed as blue triangles for Presentation 1 (left), blue circles for Presentation 2 (middle), and white circles for Presentation 3 (right).

Task 1 – Cockpit simulation flight

During a simulated flight, pilots navigated a pre-planned route from point to point while performing common pilot tasks like *see and avoid* (during VFR), reading charts, operating radio and navigational frequencies, listening to radio communications, viewing approach plates, and observing the cockpit instruments and the weather presentation. While performing these tasks, pilots typically allocate their focus of attention to distinct cockpit areas corresponding to the out-the-window view, the glass instrument display, the weather presentation, the console, and the sectional map. In the course of pilots' multitasking, we introduced METAR-symbol changes that signaled reduced ceiling and visibility conditions at selected airports. Our main interest was to see whether pilots could detect these symbol changes and whether the perception of change was the same for pilots using different weather presentations.

Study participants

Sixty instrument-rated GA pilots (56 male and 4 female) volunteered to participate in Tasks 1 and 2. In addition, Task 2 included an additional four non-instrument-rated (all male) pilots. The

participants were recruited from the pool of federally employed and contract pilots at the Federal Aviation Administration William J. Hughes Technical Center. Participants were paid their regular hourly rate while participating. Participants were randomly assigned to one of three weather presentations; each presentation depicted the same weather information but employed different symbols and colors (see Figure 2).
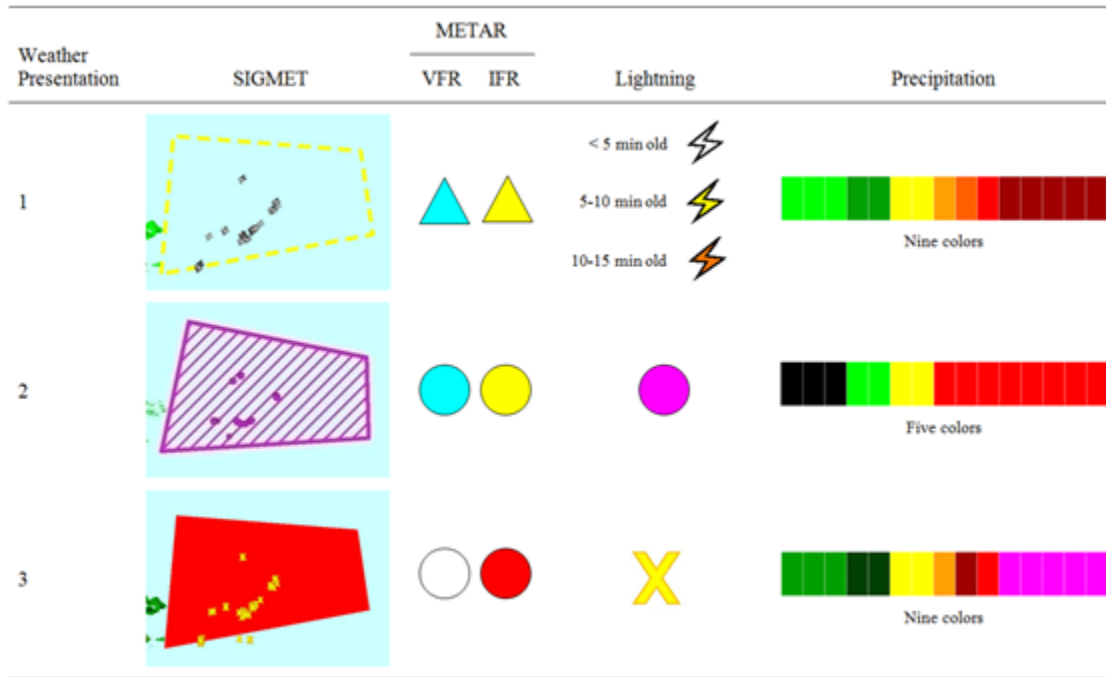
Figure 2. The three weather presentations used in the study (top: Presentation 1, middle: Presentation 2, and bottom: Presentation 3). All three weather presentations provide the same information but use different symbols and colors.

Data recordings

During the simulation flights, we recorded several data variables including the aircraft's altitude and heading. We also measured how often pilots contacted the air traffic controller to request deviations or weather updates. We also recorded if pilots detected the scripted METAR color changes that occurred at 10, 19, and 30 minutes into the flight scenario. Finally, we measured pilots' mental workload.

Results from the simulation flights

We found that overall flying behavior—as measured by altitude and heading changes—did not differ credibly between the three weather presentation symbologies. Pilot communication with the air traffic controller (i.e., frequency and duration) was also similar across the three weather presentation symbologies.

However, we found important differences in the METAR color "change-detection" accuracy between pilot groups using different weather symbologies (see Figure 3). While the detection performance for pilots using Presentation 3 was modest (62%), METAR color change-detection performance was poor for pilots using Presentation 1 (25%) and Presentation 2 (37%).
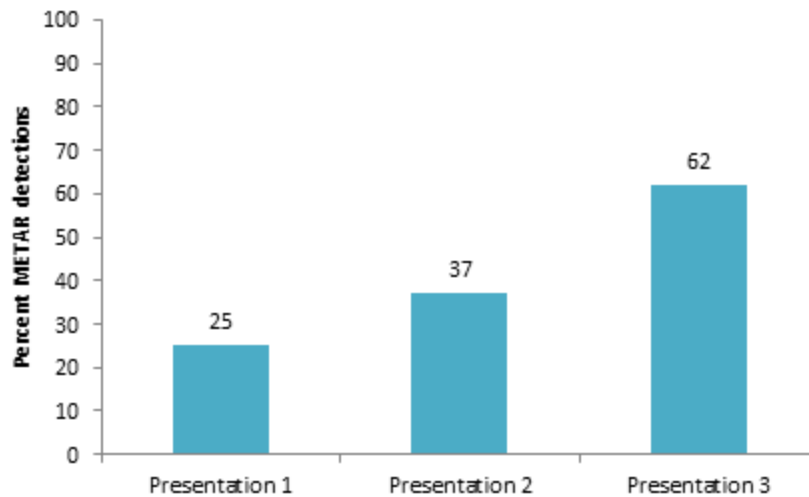
Figure 3. Overall percent METAR color change detections during flight for the three weather presentations (the data is summarized for METAR color change detections at 10, 19, and 30 minutes into the flight scenario).

We also found that pilots who detected the METAR-symbol color changes differed in their mental workload compared to pilots who did not detect color changes. In most cases, detecting a METAR color led to a temporary increase in pilots' mental workload—this was, typically, due to increased flight-planning and decision-making activity—as evidenced by pilots' requests for detailed weather information and/or decisions to change their destination airport.

Task 2 – Detection of symbol change

Task 1 addressed, specifically, pilot detection of METAR-symbol color changes during a realistic and representative piloting task. In Task 2, our aim was to measure the change-detection performance for all the weather graphics (shown in Figure 2) in a separate and isolated task.

During Task 2, each pilot viewed pairs of weather-presentation images constructed using the same symbology that they encountered in Task 1. After the first image was presented on-screen for a few seconds, the screen was blanked and then the second image appeared. We systematically manipulated one weather element (i.e., METARS, precipitation, SIGMET, lightning, and time-stamp) in one of the two weather-presentation images; for example, by showing yellow METARs in the first image and showing blue METARs in the second image or by showing the SIGMET outline in the second image but not in the first image. To better assess pilots' ability to determine whether a change occurred, we also included some trials (i.e., pairs of images) in which there were no changes to weather elements.

Participant task

After reading the on-screen instructions, participants first completed 14 practice trials followed by 60 test trials. The pilots initiated each trial by pressing the spacebar on the keyboard. They responded (by pressing the key labeled *Yes*) if they detected a change or responded (by pressing the key labeled *No*) if they did not detect a change.

Data recordings

In Task 2, we measured the change-detection accuracy (correct responses) for each weather information element and how long it took the pilots to respond.

Results from the symbol change detection

The result from the change-detection experiment shows that the detection accuracy varies greatly between different weather symbols and between different weather presentations (see Figure 4). Although the average change-detection performance is high across all weather presentations for precipitation areas (on average, 89% to 94% correct detections), SIGMET areas (83% to 93%), and METAR symbols (83% to 91%), pilots had difficulty seeing changes to lightning symbols (17% to 43%) and time-stamp information (13% to 20%).
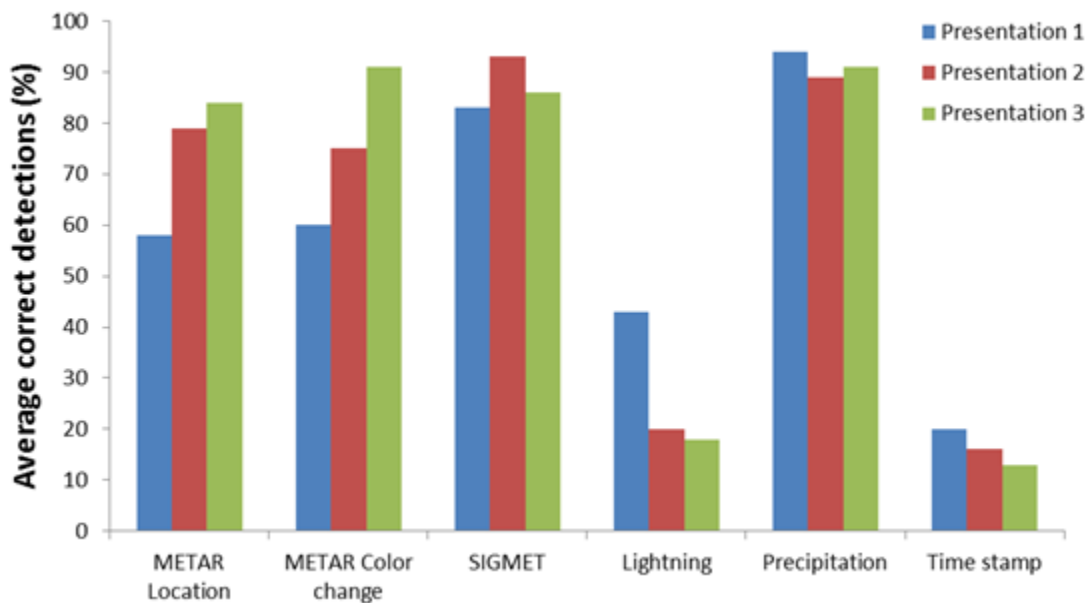
Figure 4. Average percent correct symbol detections from Task 2 for the three weather presentations.

Conclusions

This study clearly shows that weather presentation symbology affects pilots' ability to recognize symbol change and mental workload. Pilot performance varies credibly between different symbology renderings of the same weather data. Although this is a negative outcome considering the vast number of weather symbologies available, it is important empirical information that can help us examine and recommend more optimal presentations (e.g., by optimizing symbol shapes and colors).

Modern electronic cockpit displays and hand-held devices use graphical symbols to represent weather-information elements. Therefore, weather presentations should display symbols that allow rapid encoding and detection. This is especially important considering the large number of different weather elements that can be overlaid on modern multifunctional displays using different backgrounds. As more symbols and background areas are color-coded, the possible combinations of foreground and background colors rapidly increase. This can lead to salient problems where more important information (e.g., METAR-symbol color change) fails to visually segregate from less critical

background information. We need presentation symbologies that achieve good margins of legibility and detectability for all combinations of symbols and background colors.

Weather information updates during flight could, potentially, assist pilots in avoiding numerous adverse weather situations. To be effective, pilots need to perceive symbol-location and symbol-status changes to maintain their weather situational awareness. However, as symbols update their location and change colors, pilots often cannot detect the changes. Therefore, we need to continue to examine and recommend enhanced weather presentations that not only present weather elements, but that maximize pilots' ability to detect the symbol changes and their awareness of information updates with minimal cognitive processing required by the pilot.

1. INTRODUCTION

Visual Flight Rules (VFR) flight into Instrument Meteorological Conditions (IMC) is a major safety hazard for General Aviation (GA) pilots (Ison, 2014). Because of this, pilots need to assess weather information as part of their pre-flight planning. Pilots also need to maintain good "weather situation awareness" while in flight. One increasingly popular method for receiving in-flight weather updates is the use of cockpit weather displays, like certified installed display systems or the subscription to commercial weather products that can be viewed on portable devices or on cell phones (Federal Aviation Administration [FAA], 2010). These products allow the user to display a wide variety of weather information elements including precipitation, wind, lightning, echo top, and aviation routine weather reports (METARs) to mention a few. Typically, these commercial products contain both text-based and graphical presentations of weather information.

Currently, there are no industry standards for the display of weather information in the cockpit. This has resulted in large symbology variations between commercial vendors (FAA, 2010). This raises the question of whether different symbols for the same weather data have an effect on pilot perception and behavior. Ahlstrom and Dworsky (2012) found credible effects on pilot weather deviations, visual scan behavior, and cognitive engagement for different pilot groups using different weather symbology renderings. This implies that beyond studying how pilots use weather symbol information we also need to study the effects of how symbols are presented.

Although not an official standard, the RTCA provides some guidance for the display of weather data in the cockpit RTCA (2004). However, the RTCA guidance is not based on a large set of empirical data. The lack of empirical data and weather symbology standards could mean that weather symbology presentations are not optimally designed for single-pilot use while in flight. McDougall, de Bruijn, and Curry (2000) documented that well-designed symbols improve operator performance and attention management. In contrast, ill-designed weather symbols could possibly decrease usability, thereby increasing pilots' cognitive workload, degrading pilots' weather situational awareness, and impacting pilots' decision-making capability for weather-related events. Because of these detrimental effects on pilot behavior and decision-making, ill-designed weather symbols could also degrade pilot safety margins.

To support weather situation awareness, McAdaragh (2002) points out that weather displays should be intuitive, allowing pilots to act on information while navigating and piloting. That is, weather presentations should be compatible with multitasking situations that require divided attention. The weather presentation should also allow pilots to recognize weather conditions, identify individual weather symbols and differentiate them from other symbols, and support the pilot in deciding upon a course of action (Grasse, Schilke, & Schiefele, 2008). The display should support rapid interpretation and understanding, thereby reducing the cognitive resources needed to detect, analyze, and interpret graphical weather data.

1.1 Background

Previous research on the effect of weather information on pilot behavior has examined the use of graphical precipitation information. For example, Beringer and Ball (2004) compared the behavior of pilots using the Next Generation Radar (NEXRAD) information at varying levels of resolution. They found that pilots who relied more on high-resolution NEXRAD images attempted to navigate between weather more than pilots with low-resolution displays. This suggests that pilots will take higher risks going through weather systems if they are using the NEXRAD high-resolution system. This tactical use (Latorella & Chanberlain, 2002b) of NEXRAD displays is, potentially,

dangerous because there are timing issues with NEXRAD displays that create a temporal uncertainty with respect to the actual weather location. Elgin and Thomas (2004) pointed out that NEXRAD information may be 5 minutes old when it reaches the weather-service provider. It takes another minute or two for the service provider to broadcast the data. Furthermore, the cockpit display updates only once every 5−7 minutes. This process results in weather-data displays that can be more than 14 minutes old; by the time the pilot sees the information, it may no longer be accurate. This could be a serious problem if pilots use NEXRAD information as a guide when flying between hazardous weather areas.

Currently, many commercial weather products display time-stamp information that informs the pilot about the age of displayed weather data (FAA, 2010). A General Aviation (GA) pilot study by Latorella and Chamberlain (2002a) used a Graphical Weather Information System (GWIS) with a NEXRAD time-stamp. The time-stamp information was located in the upper left corner of the display, presenting the NEXRAD date and creation time. Latorella and Chamberlain found that few pilots used the time-stamp information and that pilots were uncertain about the age of the NEXRAD data. Some pilots suggested including an alert for when the NEXRAD data were too old to be used. Nevertheless, some pilots made comments that indicated they felt comfortable flying between convective weather during IMC; a potentially dangerous situation. A similar result was reported by Yuchnovicz, Novacek, Burgess, Heck, and Stokes (2001). They found that the compelling nature of NEXRAD images caused some pilots to depend too heavily on the weather display, highlighting the fact that pilots' use of a weather display does not necessarily result in optimal behavior and decision-making.

A cockpit display study by Johnson, Wiegmann, and Wickens (2006) assessed pilot use of visibility and ceiling information. Researchers included METAR symbols on their moving map display to indicate ceiling, wind, visibility, and flight category information. Potentially, pilots could use the METAR information to avoid cloud penetration. However, the researchers found a very modest effect of the METAR information, with only two pilots using the information strategically and descending before encountering deteriorating weather.

Graphical METAR information was also studied by Coyne, Baldwin and Latorella (2005) using color-coded symbols with visibility and ceiling information. In the experiment, pilots viewed movie clips of an out-the-window (OTW) scene while having access to graphical METAR symbols on a secondary display. In their experimental manipulation, the METAR symbols could either indicate better or worse conditions than the OTW. The researchers found that the METAR symbols did affect pilots' estimated ceiling and visibility values. When the METAR symbols indicated better conditions than the OTW presentation, pilots provided ceiling and visibility values that were positively biased. Researchers found the same trend when the METAR symbols indicated worse conditions than the OTW with ceiling and visibility estimates that were biased towards lower estimated values than the actual conditions.

In summary, graphical weather presentations could, potentially, aid a pilot and contribute to enhance weather situational awareness during flight. However, previous research has found both null effects and negative effects on pilot behavior and decision-making from the use of graphical weather data. In some cases the graphical weather information was present but not used; in other cases it was only used frugally. This brings up the question of legibility of different presentation symbologies (i.e., colors and shapes) and to what extent pilots are aware of symbols and can accurately perceive symbol changes. In particular, more data are needed on how pilots perceive symbol location and symbol status during the multitasking situations encountered while piloting as

symbol changes can convey important information that could impact pilot decision-making and route choice.

1.2 Purpose

The overarching goal of this study is to perform a human factors assessment of the effects of variations in cockpit weather symbology on GA pilot symbol perception. In Experiment 1, we assess pilot perception of METAR symbols during flight and how this affects flight behavior, cognitive engagement, and decision-making. In Experiment 2, we focus on pilot perception of time-stamps and weather symbols in a change-detection experiment.

2. EXPERIMENT 1

2.1 Method

2.1.1 Participants

Sixty instrument-rated GA pilots (56 male and 4 female) volunteered to participate in the study. The participants were recruited from the pool of federally employed and contract pilots at the FAA William J. Hughes Technical Center (WJHTC). Participants were paid their regular hourly rate while participating. Participants were randomly assigned to one of three Weather Presentation (WP) symbologies. Participant characteristics are described in Table 1.

Table 1. Descriptive Characteristics of Study Participants by Weather Presentation

| | | | | Flight hours accrued | | | | | |
| | | Age (years) | | Total | | Instrument | | Instrument: Last 6 months | |
Weather presentation	n	Mdn	Range	Mdn	Range	Mdn	Range	Mdn	Range
1	20	61.5	28–81	3950	225–25,000	350	10–25,000	1	0–80
2	20	65.0	30–85	5100	600–17,000	500	15–17,000	4	0–50
3	20	53.5	29–77	3450	500–35,000	400	25–35,000	4	0–100

Note. In Table 1, Mdn stands for Median (the median is the numerical value separating the upper half of a data sample from the lower half).

In comparison to FAA statistics on the age of active pilots, the participants recruited for study were generally older than the average age of student (31.5 years), sport (54.7 years), recreational (47.8 years), and private (48.3 years) pilots (FAA, 2012; http://www.faa.gov/data_research/ aviation_data_statistics/civil_airmen_statistics/2012/media/Air13-2012.xls).

2.1.2 Testing Facility

The simulation study was conducted in the Cockpit Simulation Facility at the FAA WJHTC.

2.1.3 Materials

2.1.3.1 Biographical questionnaire

The biographical questionnaire included free-response and multiple-choice questions, regarding each participant's piloting qualifications, age, accumulated flight hours, and experience with in-cockpit WPs (see Appendix A).

2.1.3.2 Flight reference materials

The participants had access to the following materials during the briefing and the simulation flight:

- Printed very high frequency (VHF) omnidirectional radio range (VOR)-to-VOR flight plan, including intersections, and VOR frequencies.

- Washington Section Aeronautical Chart (U.S. DOT/FAA, 2010) 88th edition.

- New York Section Aeronautical Chart (U.S. DOT/FAA, 2011) 84th edition (paper version).

- IFR En route Low Altitude chart – U.S., panels L29/L30 (U.S. DOT/FAA, 2011; paper version).

- Weather briefing information for the planned route, obtained from the FAA Direct Access User Terminal System (DUATS: www.duats.com; see Appendix B). The full DUATS information was condensed to two pages.

- U.S. Terminal Procedures Publication Northeast (NE) Vol. 3 of 4.

- U.S. Terminal Procedures Publication Northeast (NE) Vol. 4 of 4.

2.1.3.3 METAR-change probe questions

To ascertain whether pilots detected the METAR changes, we employed a modified version of the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995). The SAGAT involves administering a series of targeted probe questions during brief, temporary freezes in simulated scenarios. The SAGAT is administered at several time points during the scenario and is used to assess participants' ability (a) to perceive and comprehend cues and (b) to anticipate the future state of the system. As we were primarily interested in participants' ability to detect the METAR changes introduced at the 10-, 19-, and 30-minute marks, we froze the simulation at the 11-, 20-, and 35-minute marks, respectively. At each freeze point, a question designed to assess whether participants detected the METAR change (e.g., Were there any thunderstorms or other weather-related changes in the areas of Dulles and Martinsburg?) was embedded in a set of flight-related distractor questions (see Appendix C).

2.1.3.4 Post-simulation weather presentation questionnaire

The post-simulation WP questionnaire included eight items designed to elicit participants' subjective ratings of the WP (see Appendix D). Participants responded to each item using a 6-point Likert-type scale. For example, items included the degree to which participants felt the WP affected their decision-making (rating anchors: 1 = *not at all*, 6 = *very much*) and the ease with which participants were able to determine their aircraft's position (rating anchors: 1 = *very hard*, 6 = *very easy*).

2.1.4 Apparatus

2.1.4.1 Micro-Jet cockpit simulator

This study used the FAA Cockpit Simulation Facility's custom-built Micro-Jet cockpit simulator, which was configured to simulate a Mooney Bravo single-engine aircraft. The simulator is an integrated system that comprised a simulator-technician workstation, a cockpit system, and a voice communications system.

2.1.4.2 Simulator-technician workstation

The simulator-technician workstation—consisting of three desktop computers, each with its own 19-inch monitor—was located adjacent to the cockpit system. The simulator technician used the workstation to program the flight simulator and monitor its use during simulated flight. Each computer served a specific function:

- One computer, which acted as a server, ran Microsoft Flight Simulator 2004 software (Microsoft Corporation - Redmond, WA) that was used to control the flight characteristics of the simulator.

- A second computer, which acted as a client, ran Microsoft ESP software (Microsoft Corporation - Redmond, WA) and provided the visual OTW "out-the-window" display in the cockpit.

- A third computer ran (a) the G1000 Type General Aviation Glass Cockpit software (www.projectmagenta.com) to display the aircraft's control scheme (see Figure 1, Figure 2, and Figure 3); (b) the GA WP software (AeroTech Research - Newport News, VA) to display the in-cockpit WP; (c) the in-house, custom-designed data collection software to record simulator variables (e.g., altitude, bearing) at a frequency of 1 Hz; (d) the Plexcomm Virtual Radio software (Plexsys Interface Products - Camas, WA) that functioned as the simulated aircraft's radio, and (e) an audio recorder to record the pilots' verbalizations (i.e., radio communications, responses to probe questions).

2.1.4.3 Cockpit system

The cockpit system comprised a physical, two-seat, side-by-side cockpit shell and an iGATE Mod Works Smart Panel instrument panel and aircraft controls (Elite Simulation Solutions - Ovideo, FL), with a 19-inch flat-screen monitor to display the aircraft's control scheme and the track-up-configuration WP (see Figure 1 and Figure 2). A button next to the monitor allowed pilots to toggle between three zoom levels (i.e., 5, 20, or 50 nautical mile per ring) during simulated flight. An Ostendo CRVD® 43-inch curved monitor (Ostendo Technologies - Carlsbad, CA) displayed the simulated OTW view (see Figure 3).

Figure 1. The aircraft's control scheme and the track-up-configuration weather presentation.
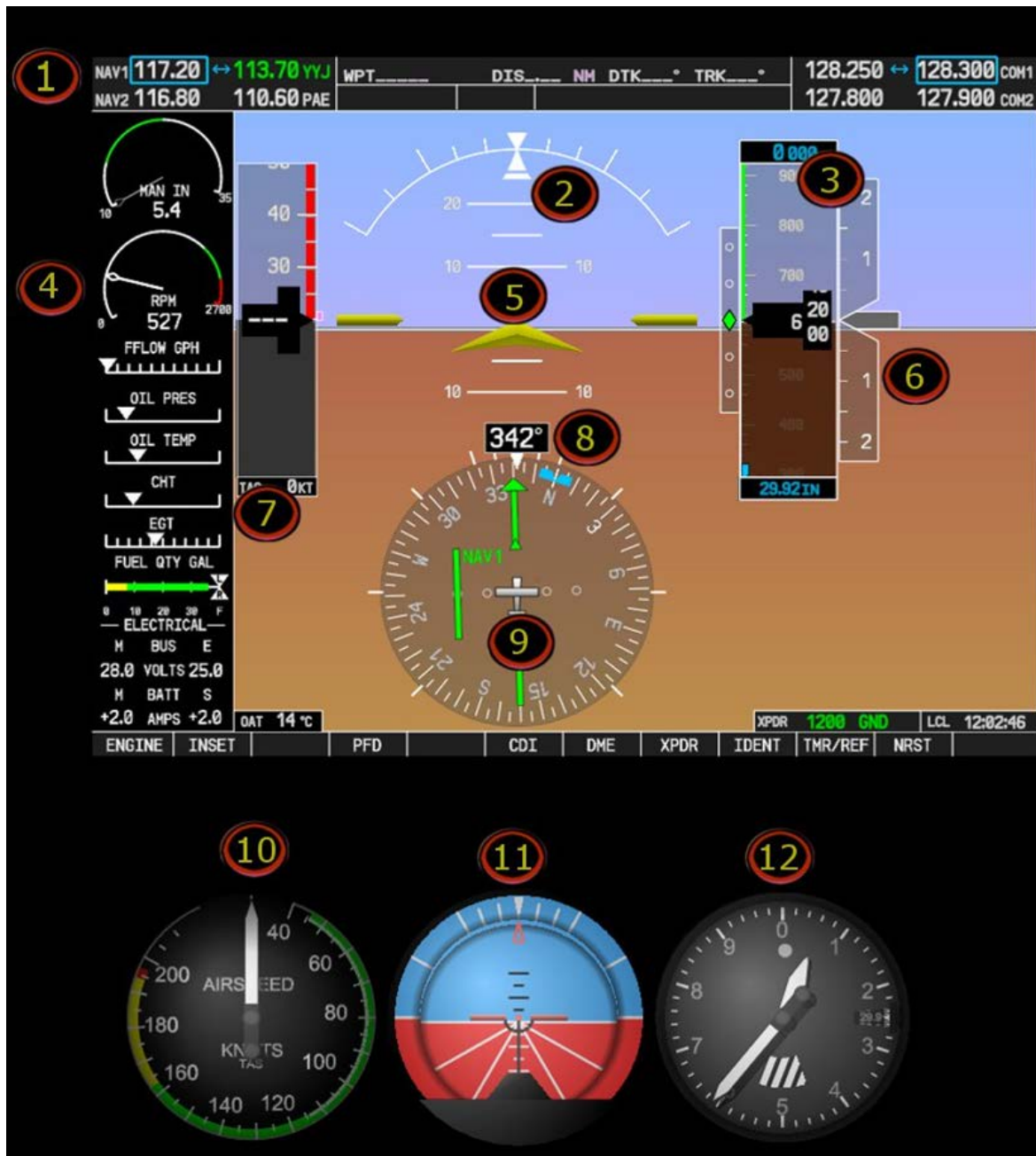
1) NAV radio
2) Turn Coordinator
3) Altimeter
4) RPM Gauge
5) Attitude Indicator
6) Vertical Speed Indicator
7) Airspeed Indicator
8) Heading
9) HIS (Horizontal Situation Indicator)
10) Backup Airspeed Indicator
11) Backup Attitude Indicator
12) Backup Altimeter

Figure 2. Project Magenta's GA glass cockpit software control scheme and element definition (bottom).

Figure 3. Micro-Jet cockpit simulator.

2.1.4.4 Voice communications systems

The voice communication system provided a link between the participant pilot in the cockpit simulator and an Air Traffic Control subject-matter expert (ATC SME), who was located in an isolated *air-traffic-control-center room* during the simulation. The system recorded the times, durations, and content of pilot-ATC voice communications for subsequent analysis. Inside the cockpit, the wired, two-way voice communication system consisted of a push-to-talk (PTT) button mounted on the control yoke and a headset worn by the participant. In the air-traffic-control-center room, the system comprised three computers—all running Plexcomm Virtual Radio software (Plexsys Interface Products - Camas, WA):

- The primary computer facilitated two-way communication between the ATC and the participant pilot, as per standard flight operations.

- A secondary computer, which was a duplicate of the primary computer, served as a backup radio that the ATC could use in case the primary computer malfunctioned.

- A third computer played a continuous, pre-recorded audio track of en-route, ATC communications to simulate background radio traffic during the simulation.

8

2.1.4.5 Functional near-infrared spectroscopy

To assess participants' cognitive engagement during the simulation, we recorded prefrontal cortical activity using a continuous-wave Functional Near-Infrared Spectroscopy (fNIR) system (fNIR Devices Model 1100 - Potomac, MD). The fNIR system uses specific wavelengths of light to measure changes in the relative ratios of oxygenated and deoxygenated hemoglobin during brain activity. The fNIR system, which was located adjacent to the simulator technician's workstation, comprised a

- wired, flexible forehead sensor pad that contained four light sources (peak wavelengths at 730 nm and 850 nm) and 10 detectors. This sensor configuration generated a total of 16 measurement locations, or voxels, per wavelength. With two wavelengths and dark current recordings for each of the 16 voxels, the system generated a total of 48 measurements for each 2 Hz sampling period (Izzetoglu, Bunce, Izzetoglu, Onaral, & Pourrezaei, 2007; Izzetoglu, Bunce, Shewokis, & Ayaz, 2010);

- control unit with integrated power supply; and

- computer, which was used to calibrate the sensor and store recorded data.

2.1.4.6 Flight plan

The flight was planned to depart from KABE (Allentown, Pennsylvania) and land at KMRB (Martinsburg, West Virginia). The flight plan followed a VOR-to-VOR route from KABE to ETX (East Texas VOR), LRP (Lancaster VOR), VINNY intersection, EMI (Westminster VOR), MRB (Martinsburg VOR), and then to KMRB (see Figure 4). As the goal of the study was to investigate the effect of the WP on pilot cognitive engagement and decision-making during at-altitude flight, we opted to maximize the time spent in at-altitude flight by omitting the take-off and landing phases. The simulation, therefore, started with the aircraft at a cruising altitude of 8,500 ft in the area of KABE, headed toward ETX; the local time was 2:00 PM on June 9, 2011.

Figure 4. Simulation scenario route from Allentown to Martinsburg airport (KMRB).

2.1.4.7 Weather-related changes

The main goal of the study was to evaluate the effect of WP symbology on pilot behavior, decision-making, and cognitive engagement. Specifically, we were interested in the effect of weather symbology on pilots' ability to detect changes in METAR status from VFR to IFR conditions. At the start of the scenario, all METAR symbols along the route indicated VFR conditions. To assess pilots' ability to detect changes in METAR status, selected METARs in the region of the planned flight route were programmed to change from VFR conditions to IFR conditions at three time points during the 35-minute simulated flight. The first METAR change occurred at 10 minutes into the scenario, the second at 19 minutes, and the third and last METAR change occurred at 30 minutes (see Figure 5). Table 2 describes the timing of the changes, and the METAR(s) that changed at each point. After a METAR changed from VFR to IFR status, it remained IFR for the remainder of the simulation.

| t = 0 minutes | t = 10 minutes | t = 19 minutes | t = 30 minutes |
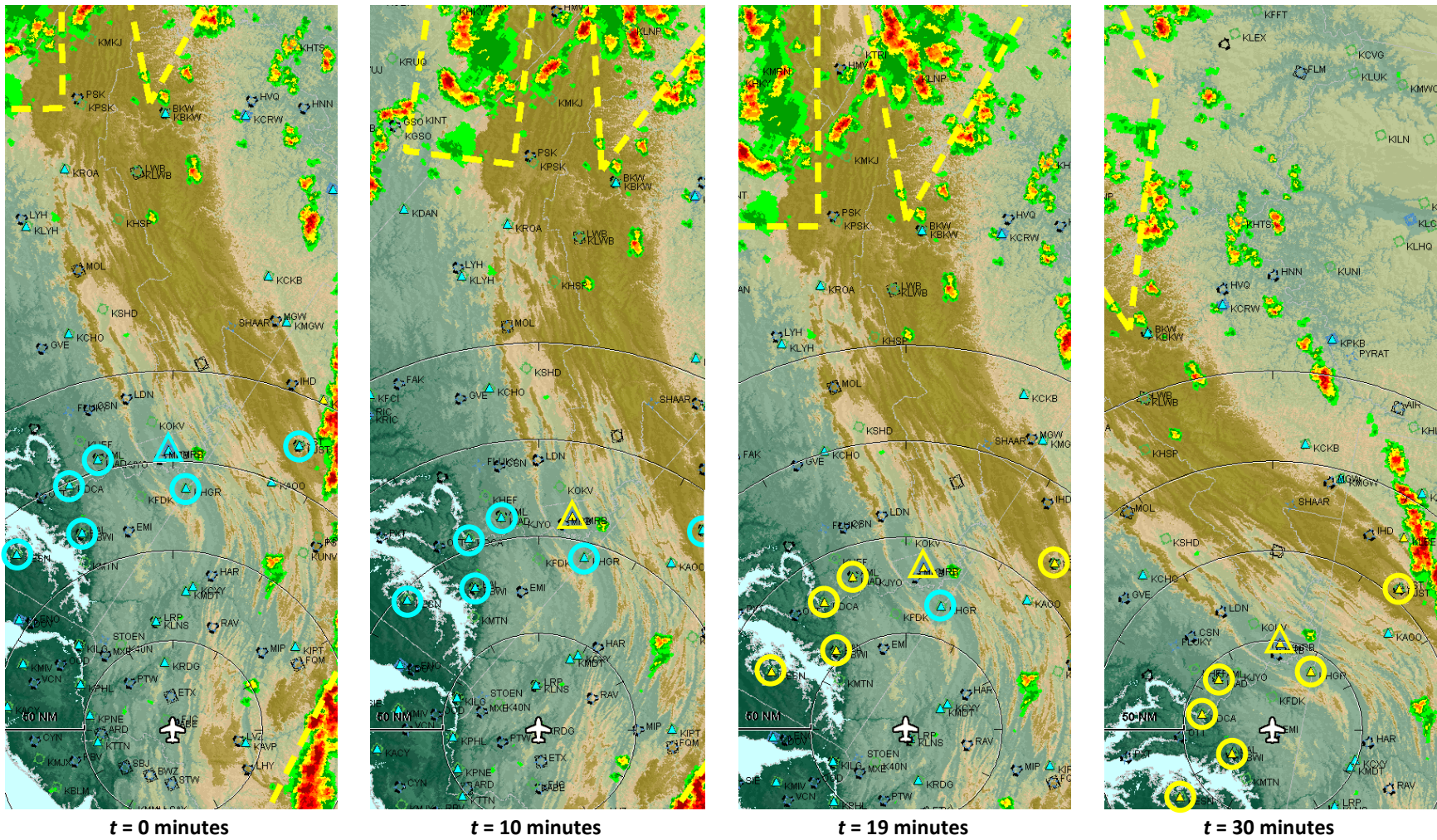
Figure 5. WPs showing initial (t = 0 minutes) VFR state of all METARs in the area of the planned flight. The METARs that change from VFR to IFR during the simulated flights are highlighted (Δ = destination airport, Ο = six remaining METARs). At t = 10 minutes, the METAR at the destination airport changes to IFR. At t = 19 minutes, five other METARs changes from VFR to IFR. At t = 30 minutes, the seventh and last VFR METAR changes to IFR. Note: This is presented here using WP 1.

Table 2. METAR Changes

Change #	Scenario time (minutes)	Number of METARS changed	METAR(s) Changed from VFR to IFR	
			METAR code	Location
1	10	1	KMRB	Martinsburg, WV
2	19	5	KBWI	Baltimore/Washington International, MD
			KDCA	Washington National, DC
			KESN	Easton/Newman, MD
			KIAD	Washington, DC/Dulles, VA
			KJST	Johnstown, PA
3	30	1	KHGS	Hagerstown, MD

2.1.5 Weather Information

The cockpit WPs incorporated four types of weather information, which were overlaid on an active map (see Figure 6).

- **Aviation routine weather report (METAR) for specific locations.** Small, color-coded symbols were used to summarize each METAR as either Visual Flight Rules (VFR) or Instrument Flight Rules (IFR) flight conditions, according to visibility and ceiling; marginal VFR and marginal IFR conditions were not included in this study.

- **Significant Meteorological Advisory (SIGMET) information.** Depicts advisories on weather that is significant to the safety of all aircraft. Polygons (i.e., rectangles)—formed by continuous or dashed lines—marked the regions affected by SIGMETs.

- **Lightning strikes.** Small symbols marked the regions affected by lightning strikes.

- **Precipitation.** Color-coded shading depicted the intensity of precipitation.
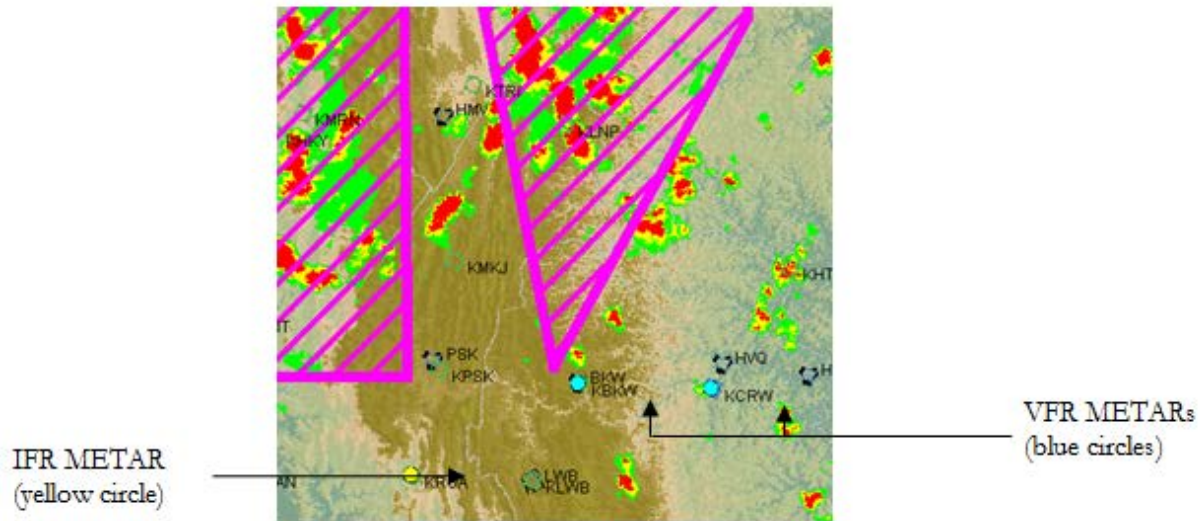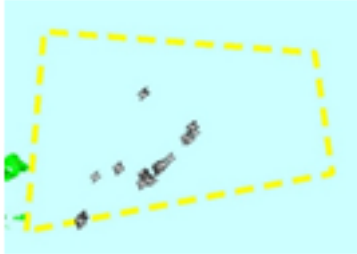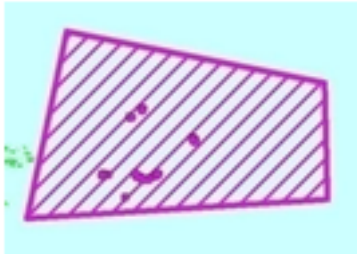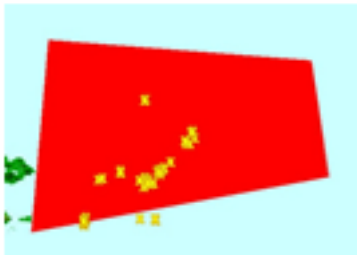
Figure 6. Portion of a weather presentation, showing the different weather-information types.

2.1.6 Weather Presentation Symbology

Following on from previous research that compared the effect of WP symbology on weather avoidance during flight (Ahlstrom & Dworsky, 2012), we compared three WP symbologies that varied in the symbols (shapes and colors) used to represent the weather information (see Table 3). The symbologies are representative of current industry implementations (i.e., existing presentation symbology used today in commercial products). Figure 7 illustrates a sample of weather data presented using each of the three display variations.

Table 3. Weather Presentations (WPs)

Presentation	SIGMET	METAR		Lightning	Precipitation
		VFR	IFR		
1				< 5 min old / 5-10 min old / 10-15 min old	Nine colors
2					Five colors
3					Nine colors
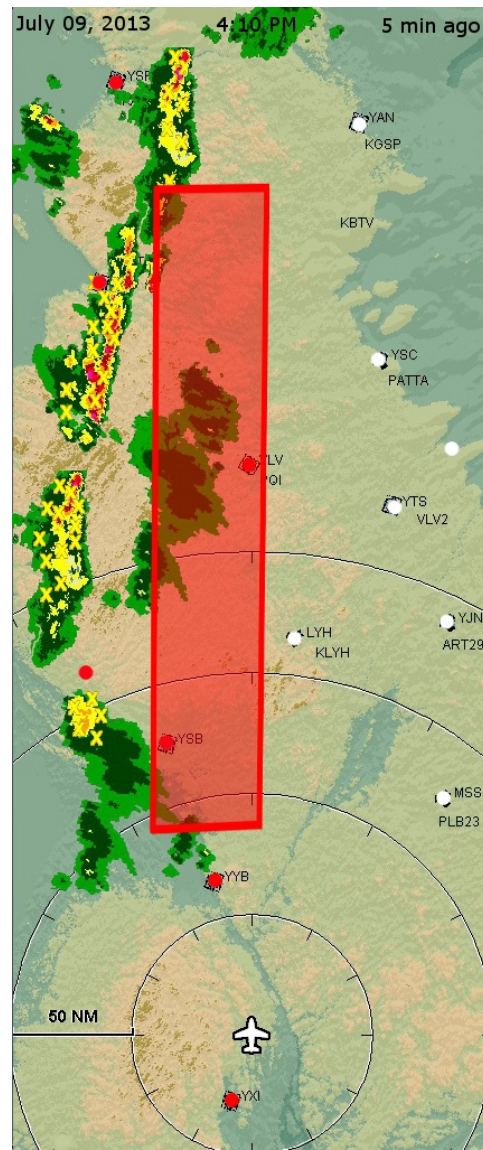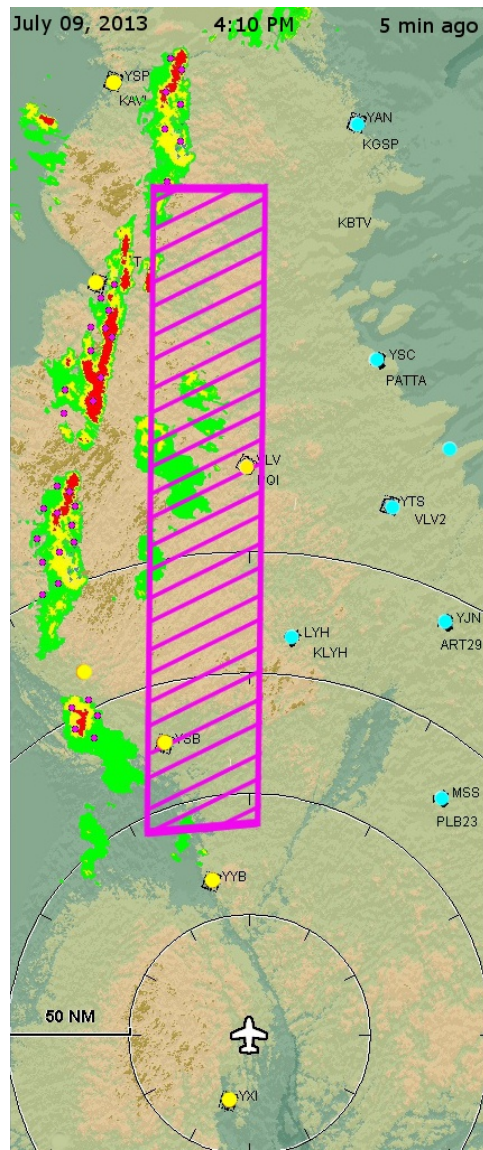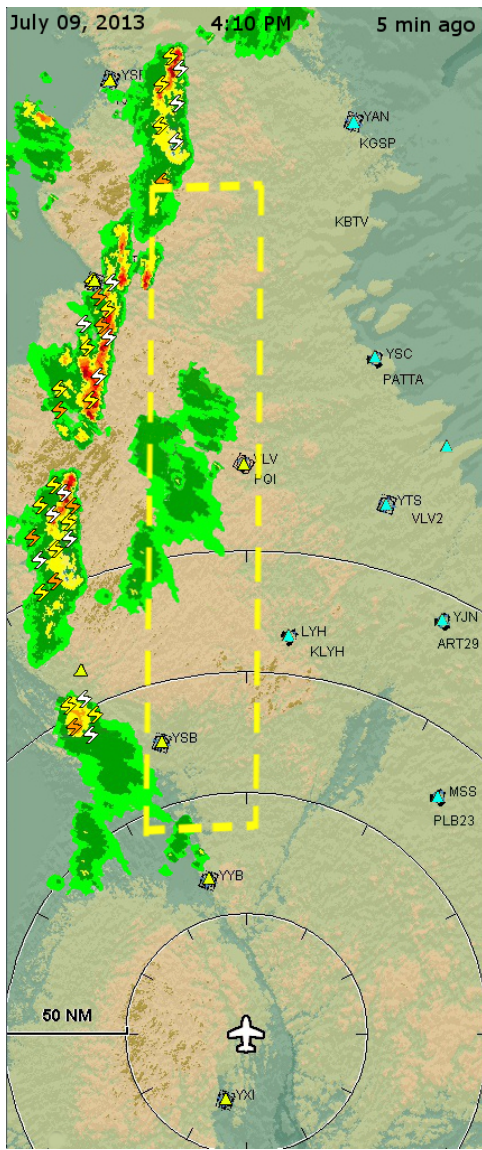
Figure 7. A sample of weather data presented using the three weather presentations (WPs).

2.2 Procedure

Upon arriving at the test facility, participants read and signed an informed consent form. Next, the ATC SME (who was also a qualified pilot) briefed the participant about the (a) flight plan; (b) weather information; (c) aircraft controls and functions (i.e., navigation, autopilot, and horizontal situation indicator [HSI]); and (d) fNIR system. The briefing took place in a briefing room, and was presented as a Microsoft PowerPoint slideshow. During the briefing, the ATC instructed participants to assume that they:

- Were an IFR-rated pilot, but had chosen to fly using VFR.

- Have two passengers on board.

- Have an important business meeting at the destination airport.

- Had planned the route themselves the previous day.

- Had previously flown the planned route using IFR, but had chosen—today—to fly using VFR.

- Had, prior to take-off, established communications with Allentown Approach and are receiving flight following.

Following the briefing, participants were given unlimited time to peruse the flight reference materials. The ATC SME then escorted the participant to the Micro-Jet cockpit simulator for a 15-minute practice flight, during which the ATC familiarized the participant with the simulator's (a) aircraft controls, (b) WP and zoom function, (c) radio, and (d) HSI. Following the practice flight, the ATC moved to the isolated ATC Center room and prepared to assume the role of air traffic controller, while the fNIR technician fitted and calibrated the fNIR sensor. When the pilots indicated that they were ready to begin the simulation flight, the simulator technician started the simulation; simultaneously, the fNIR technician initiated the fNIR-signal recording.

At the predetermined freeze points (i.e., 11, 20, 35 minutes) the simulator technician froze (i.e., paused) the simulation and reactivated the audio recorder. The primary researcher covered the control scheme/WP screen to prevent participants from referring to the displayed information and then presented participants with a printed copy of the probe questions (see Appendix C for the specific questions at each freeze point). The researcher instructed participants to read and answer each question out loud. After participants answered the probe questions at the 11- and 20-minute freeze points, the researcher uncovered the control scheme/WP screen and the simulator technician resumed the simulation. After participants answered the probe questions at the 35-minute freeze point, the simulation technician ended the simulation.

At the conclusion of the simulation flight, participants returned to the briefing room and completed the weather-presentation questionnaire. The primary researcher then handed-off the participant to the researcher running Experiment 2.

2.2.1 Independent Variable: Weather Presentation

We manipulated the independent variable WP by presenting weather information under three different symbology modes. In the following, we refer to these three WPs as 1, 2, and 3. The data types that we are using in the present simulation presentations are Precipitation, Meteorological Report (METAR), Significant Meteorological Advisory (SIGMET), and Lightning (as shown in Table 3 and Figure 7). For Experiment 1, we will focus specifically on in-flight changes to flight category information (i.e., change from VFR to IFR parameters; see Table 4).

Table 4. METAR Weather Information Symbology for VFR and IFR Flight Parameters Across the Three WPs

Flight parameters	WP		
	1	2	3
VFR	▲ (cyan triangle)	● (cyan circle)	○ (white circle)
IFR	▲ (yellow triangle)	● (yellow circle)	● (red circle)

2.2.2 Description of Weather-Information Types

2.2.2.1 METAR variations

Our simulation WPs contain aviation routine weather report (METAR) symbols that can be broken down into four weather parameters that are based on the flight category (see Table 5). For the present study, however, we used only METAR symbols for VFR and IFR.

Table 5. Flight Categories

Category	Ceiling		Visibility
Low Instrument Flight Rules (LIFR)	< 500 feet AGL	And/Or	< 1 mile
Instrument Flight Rules (IFR)	500 to 1,000 feet AGL	And/Or	1 mile to 3 miles
Marginal Visual Flight Rules (MVFR)	1,000 to 3,000 feet AGL	And/Or	3 to 5 miles
Visual Flight Rules (VFR)	> 3,000 feet AGL	And/Or	> 5 miles

Note. Table adapted from "Aviation Weather Services AC 00-45" by FAA & NOAA, 2010.

The three WPs have different colors and shapes to show METAR information (as previously shown in Table 4). The presentations for Variation 2 and 3 use filled circles to show flight categories; Variation 1 uses filled triangles.

2.2.2.2 Precipitation variations

Precipitation based on radar information depicts the intensity of precipitation overlaid on the active map. This data updates every 5 minutes, on average. Each of the WPs differs on the number of color codes for intensity. WP 1 and WP 3 both display nine colors for precipitation intensities; WP 2 uses five colors (as shown in Table 3).

2.2.2.3 SIGMET variations

The SIGMET information depicts advisories on weather that is significant to the safety of all aircraft. These advisories are divided into two different categories: non-convective and convective. Non-convective SIGMETs depict severe and extreme turbulence, severe icing, widespread dust, sandstorms, or volcanic ash that reduces visibility to less than 3 miles (FAA, 2014). Convective SIGMETs are issued for tornadoes, areas of thunderstorms, and hail. The SIGMET information updates every 4 hours unless a hurricane is present, in which case they are updated every 6 hours. Convective SIGMETs are updated hourly. However, our presentation updates all information every 5 minutes regardless of new information. Each of the three presentations depicted the SIGMET in different ways. WP 1 use a dashed yellow line, WP 2 showed a solid magenta outline (filled with magenta hash marks), and WP 3 showed a solid red outline (filled with red).

2.2.2.4 Lightning variations

If available, lightning strike information can help the pilot be better aware and provide better situational awareness of convective activity in the area where they are flying. All three variations present lightning information in different ways. WP 1 presented lightning information by a lightning bolt symbol, WP 2 used a magenta dot, and WP 3 used a yellow *X*.

2.2.3 Dependent Variables

During this cockpit simulation, researchers recorded dependent variables to evaluate pilot sensitivity to METAR color changes during flight, and whether detection sensitivity was affected by WP symbology. In addition, we measured how the detection of METAR changes affected pilot decision-making and flight behavior. The dependent variables capture the following categories: System Performance (aircraft and instrument panel data), Communication (pilot/ATC PTT), Weather Situation Awareness (detection of METAR changes), Decision Making (e.g., whether the pilot continues with VFR or IFR flight after METAR changes), WP Usage (zoom usage), and Cognitive Engagement (i.e., fNIR oxygenation changes). In Table 6, we provide a list of the dependent variables and a short description.

Table 6. Dependent Variable List

Number	Dependent variable	Description
1	System performance measures	Readings from instrument panels and data from the cockpit simulator.
2	Number and duration of ATC communications	The number and duration of pilot/ATC communications.
3	Weather situation awareness	SAGAT query of the detection of METAR color changes - the number of times pilots detected a METAR color change (at 11 min and 16 min into the scenario).
4	Decision-making	Pilot decision to use VFR versus IFR flight after METAR changes.
5	WP zoom	The number of zoom changes and the display duration at each of the three zoom levels.
6	Cognitive engagement	The oxygenation changes captured by the fNIR system (the system analyzes increases and decreases of oxygenated hemoglobin, which correlates with changes in cognitive engagement).

Note. ATC = Air Traffic Control; SAGAT = Situation Awareness Global Assessment Technique; METAR = Aviation Routine Weather Reports; VFR = Visual Flight Rules; WP = Weather Presentation; fNIR = Functional Near-Infrared Spectroscopy.

2.2.3.1 System performance measures

During the simulation flights, we recorded two parameters from the cockpit simulator system (Flight Simulator, 2004). We used these parameters to calculate two dependent variables that are associated with pilot flight behavior:

1. **Altitude** - The height of the airplane above mean sea level as displayed on the altimeter.

2. **Heading** - The direction which the airplane is pointed.

Researchers will use these variables to assess whether pilot flight behavior is comparable for all three WPs.

2.2.3.2 ATC communications

We recorded the PTT communication between the pilot and the controller (ATC SME) to evaluate the number and duration of radio communications. From the recordings we extracted when pilots made requests from ATC for (a) weather updates, (b) route deviations, and (c) IFR flight plans.

2.2.3.3 Weather situation awareness

We used the participant's response to the METAR probe question to determine whether the pilot had seen the METAR change or not. The pilot's responses to the SAGAT distractor queries were discarded.

2.2.3.4 Decision making

The first METAR change (from VFR to IFR) occurred 10 min into the scenario, and the second METAR change occurred 19 min into the scenario. If the pilot detected the METAR changes, the pilot could still continue with VFR—or alternatively, the pilot could contact ATC and

request an IFR flight plan. Alternatively, pilots could elect to go to one of several alternate destination airports. We were specifically interested in assessing whether—after each METAR change—pilots differed systematically in their decision-making based on their specific WP (i.e., presentation 1, 2, or 3).

2.2.3.5 WP zoom changes

During the simulation, pilots had the ability to zoom in on weather areas or routes via a button next to the presentation panel. Three zoom level settings (i.e., 5, 20, or 50 nautical mile per range ring) were available. By recording the zoom variable we can compare when pilots changed zoom levels during the scenario, which zoom level was used the most, and whether the zoom usage differed across the three WPs.

2.2.3.6 Cognitive engagement: Functional near infra-red spectroscopy

During each simulation run, we used the 16 fNIR channels to record prefrontal oxygenation changes.

2.2.4 Analysis Framework

Traditionally, researchers use the Null Hypothesis Significance Testing (NHST) framework to plan research and to analyze their study outcomes using p-values. However, the NHST framework has received a broad range of criticism (e.g., Gigerenzer, 2004; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). The core issue with NHST is that there is no single and unique p-value for any given data set; all data sets have many different p-values because the p-value is determined by the data generating procedure and the experimenter's intentions (i.e., the number of planned tests: Kruschke, 2010; Wagenmakers, 2007).

In NHST data analysis, the p-value is used as a measure of evidence in the data against a null hypothesis (H_0) of *no effect*. The logic is that the smaller the p-value, the stronger the evidence against the null hypothesis in the data. If the p-value is less than the conventional 5% significance level (e.g., $p = .025$), the researcher rejects the null hypothesis and declares a significant result. If the p-value is larger than the conventional 5% significance level (e.g., $p = .061$), the researcher declares no effect. It is also common to see the p-value used as a proxy for effect size. Study outcomes are often referred to as *significant* or even *highly significant*, rather than *statistically significant*. As Goodman (1992) notes, if the p-value is small enough to be significant, the effect is often interpreted to be *real*. This leads some researchers to also falsely interpret the result as the complement of the p-value, $1 - p$ (e.g., $1 - .05 = .95$) and to declare that the outcome would hold up in future replications with odds of 95 to 100. Researchers also commonly interpret p-values using variations of the following statements:

1. Our significance test at $p < .05$ simply means that our result would have occurred solely by chance less than 5 times in 100 significance tests.

2. In our analysis, there is a 5% probability that a significant result was due to chance when we are using a criteria of $p \leq .05$, and the probability of finding a significant result by chance increases to 20% when we are using a criteria of $p \leq .20$.

3. The alpha level indicates the rate at which our results would be expected to occur by chance, rather than real differences between our experimental conditions.

Tragically, these three statements are all false, and they are not just inconsequential statistical declarations. Rather, these statements represent a fundamental misunderstanding of the underlying core of the entire NHST inferential approach. The statements, or any derivative thereof, are referred to by Carver (1978) as the odds-against-chance fantasy. This includes any interpretation of the *p*-value as a probability that *the result is due to chance*, or *caused by chance*. Because the *p*-value is derived under the strict, up-front assumption that H_0 is true (i.e., all differences are entirely due to sampling error, the effect size equals zero with 100% probability), it is impossible for the *p*-value to be a measure of the *odds of chance*. All fantasies aside, the correct interpretation of a NHST *p*-value is:

> The probability (e.g., $p = .02$) of getting the outcome at hand, or more extreme values, given that the null hypothesis is true, $p(D|H_0)$ - contingent on following the strict a priori assumptions of the use of a particular sampling procedure, the particular set of outcomes to test, and with the assumption that no other statistical tests will ever be performed on the same data set again. (see Gigerenzer, 2004; Goodman, 2008; Hubbard & Bayarri, 2003)

As stated above, the core problem with the computation of *p*-values is that they depend entirely on the data generating procedure and the intentions of the experimenter, because these factors determine the sampling space and the derived sampling distribution. The sampling distribution, in turn, determines the *p*-value. Because of this, *there is no single and unique p-value for any given data set*. Using the same data set with different experimenter intentions (i.e., stop rules) can lead to outcomes in which *p* is both less than, and not less than, .05. Therefore, the classification of research outcomes as *significant* (if $p < .05$) or *nonsignificant* (if $p > .05$) is a flawed decision rule and a meaningless yardstick when measuring effects from study outcomes.

2.2.4.1 Bayesian estimation

An alternative to NHST and *p*-values is Bayesian estimation (Wagenmakers, 2007). The Bayesian framework provides richer and more complete information regarding data parameters, and it avoids NHST constraints as the demand for multiple test corrections and the taboo of accepting null values (Kruschke, 2011). NHST analyses provide the probability of the data values given the truth of an a priori specified null hypothesis, $p(\text{Data values}|H_0)$. But what we really want to know is the probability of parameters and model structure given our data values, $p(\text{Parameter values and model structure}|\text{Data values})$. This goal can only be accomplished by using Bayesian inference.

For Bayesian parameter estimation when Bayes' rule is applied to parameters (θ) and data (D), we have:

$$\text{(1)}$$
$$p(\theta|D) = p(D|\theta)\,p(\theta)\,/\,p(D),$$

where the posterior distribution, $p(\theta|D)$, is the result of the likelihood, $p(D|\theta)$, multiplied by the prior, $p(\theta)$, divided by the evidence, $p(D)$. The posterior is our strength of belief in the parameter values and model structure after the data are taken into account. The likelihood is the probability that the data could be generated by the model with parameter values θ. The prior is the strength of our belief in θ before we have taken the data into account. The evidence is the probability of the data according to our model—from summing across all model parameter values weighted by our strength of belief in those parameter values.

When using real-world data and Bayes' rule, one often has to compute a difficult integral in the denominator—that is, the evidence, $p(D)$—or find a suitable approximation. Fortunately, modern sampling methods, referred to as Markov Chain Monte Carlo (MCMC) methods, are available for numeric approximation of probability distributions. To analyze the data from the current experiments, we use MCMC sampling to get a good description of the posterior distribution using JAGS (Plummer, 2003, 2011) called from R (R Development Core Team, 2011) via the package rjags, using adapted program code from Kruschke (2011). This procedure involves generating a large number of representative combinations of parameter values from the posterior distribution, and then using those values to generate an approximation of the posterior.

2.2.4.2 The posterior distribution

The Bayesian analysis yields a complete distribution of credible values in the posterior distribution. Once we have a large sample of representative parameter values, we can evaluate, for example, the mean of a parameter distribution, its shape, or the difference between values of different parameters.

In the present study, we use a separate decision rule to convert the posterior distribution to a specific conclusion about a parameter value. In Figure 8, the black horizontal bar represents the 95% High Density Interval (HDI). Every value inside the HDI has a higher probability density (i.e., credibility) compared to values that are outside the HDI. Therefore, values contained within the 95% HDI represent the most credible values of the parameter. When we explore differences between parameter values in a contrast, we compute these differences at each step in the MCMC chain and plot the differences along with the HDI in a histogram. Posterior histograms show, at the same time, what differences are credible and the uncertainty of those differences. If the value 0 (implying zero difference between parameters) is not contained within a 95% HDI for a histogram of differences, we say that the difference is credible. If, on the other hand, the 95% HDI includes the value 0, we cannot say that the difference between parameter values is credible because a difference of 0 is indeed among the possible outcomes. As shown in Figure 6, using mean group accuracy, μ, the posterior difference between $\mu1 - \mu2$ is credible because the value 0 is not included in the 95% HDI. On the other hand, the posterior for the difference between $\mu3 - \mu2$ contains the value 0 within the 95% HDI, and therefore there is no credible difference between $\mu3$ and $\mu2$. The 95% HDI provides both a summary of the distribution and is a decision tool to determine what parameter values are credible.
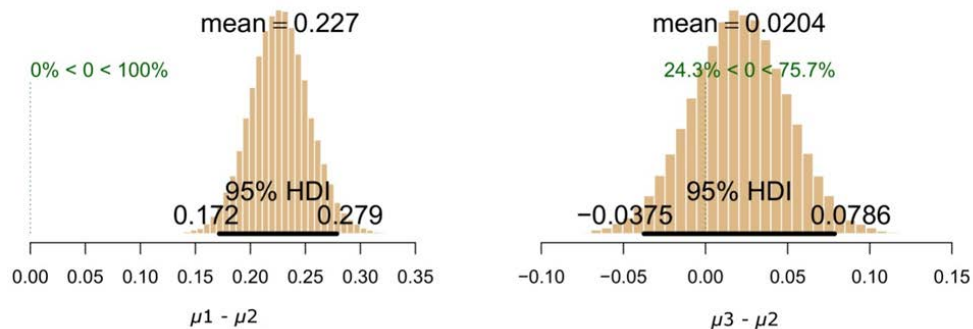
Figure 8. Histograms of posterior differences between hypothetical group means $\mu1$ and $\mu2$ (left), and $\mu3$ and $\mu2$ (right). The black horizontal bar represents the 95% HDI. The vertical dotted axis at 0.00 shows the proportion of the posterior distribution that is below and above the value 0 (i.e., 0% < 0 < 100% for the left distribution and 24.3% < 0 <75.7% for the right distribution).

2.2.4.3 The models

In the present study, one of our main interests is to assess participants' ability to detect changes to METAR symbols (i.e., the presence or absence of METAR symbols and METAR color changes) in weather presentations 1-3 (hereafter abbreviated WPs 1-3). Experiments 1 and 2 assess participants' ability to detect implicit and explicit changes, respectively. In each case, the participant's response can either be correct or incorrect, giving each participant a score of total correct responses for each experimental condition. We label each group's mean accuracy μ (i.e., *μ1, μ2,* and *μ3*), and our main interest is to assess differences in μ between the three groups for all experimental conditions. To analyze these data, we use a hierarchical model that assumes that each participant's response depends on the value of the bias parameter θ in a binomial distribution.

In Experiment 1, we are also analyzing metric predicted variables (e.g., altitude) with one nominal predictor (one-way; e.g., WP) or two nominal predictors (two-way; e.g., METAR Change and WP) using Bayesian analysis of variance (BANOVA).

Unlike NHST ANOVA, with BANOVA there is no requirement of equal variances (i.e., homogeneity of variance) for all factor levels. Instead of estimating the variance within levels of the predictor by a precision (i.e., reciprocal of the variance) that is assumed to be homogeneous across groups, we use one-way and two-way BANOVA models that accommodate non-homogeneous variances. That means that each group or condition in the model is allowed to have its own variance. When performing multiple comparisons using a Bayesian analysis, there is no need to make corrections for multiple tests (like NHST) because for each additional test we only view the outcome from different perspectives in the multidimensional parameter space. There is only one posterior distribution; it does not change as we perform multiple tests on the same data set. For a detailed description of the binomial and the BANOVA models see Kruschke (2011).

For some analyses, we also use frequency count data (e.g., number of zoom level transitions). These data values are not on a continuous metric scale, but instead fall at discrete levels and therefore should not be analyzed using models with a normal likelihood function. For these data sets, we perform Bayesian contingency table analysis using Stan (The Stan Development Team, 2013).

For all analyses, we used 200,000 samples to derive the posterior distribution. For the binomial analyses, we used 1,000 steps to *tune* the samplers and 2,000 steps to *burn-in* the samplers, while running 3 chains and saving every step in the chain (i.e., we used no *thinning*). For the BANOVA analyses, we used 1,000 steps to tune the samplers and 5,000 steps to burn-in the samplers, while running 3 chains and saving every step in the chain. For all analyses, we use priors that are vague and noncommittal on the scale of the data.

2.2.4.4 Distributions, outliers, and robust estimations

Recorded data from human-in-the-loop simulations often have distributions that are non-normal or skewed. This is a particular problem when we use normal-likelihood models. The data should, at the very least, roughly approximate a normal distribution. The more the data deviate from the normal, the worse the normal-likelihood model serves as a realistic descriptive model of the data. One way to address this problem is to logarithmically transform the data to make it more normal. In this study, we use the "base-10" logarithm transformation. This transformation is monotonic—that is, it only rescales the data while preserving order.

It is also quite common for human-in-the-loop simulation data to include data points that are far away from central data points. The issue with such data points is whether they are genuine data points or whether they are data points with unusually low or high values that are caused by irrelevant factors. These data points are often termed *outliers* and researchers have used various methods to deal with this problem (Cousineau & Chartier, 2010). One option is to decide upon some cut-off values and then reject each value in the data set if they fall outside those cut-off values. For example, Tukey proposed that if a data point is below the lower quartile – 1.5 times the interquartile range, or above the upper quartile + 1.5 times the interquartile range, it should be considered an outlier and rejected (Tukey, 1977). The main issue with rejecting data values based on cut-off values is that we often don't know whether these values are genuine data points or are noise data. Therefore, when we reject data values because they are deemed outliers we risk throwing out the baby with the bath water.

An alternative strategy is to use robust estimation. Robust estimation is called robust because it can accommodate outliers in the data. In this study, we accomplish this by using BANOVA models with *t* distributions rather than normal distributions. A *t* distribution with tall-tails (small *df*) can *reach out* and accommodate values that lie outside the central data. A normal distribution, on the other hand, will increase the standard deviation and (incorrectly) *move* the mean of the distribution towards the outlier.

2.2.4.5 Bayesian power analysis

Statistical power is the probability that researchers will accomplish their stated research goals. Modern Bayesian methods provide a logical and consistent framework for how to define and compute statistical power (Kruschke, 2011).

In this study, the overarching goal is to assess the effects of variations in cockpit weather symbology on GA pilot behavior, cognitive engagement, and decision-making. In a previous study (Ahlstrom & Dworsky, 2012), we found an effect of variations in WP symbology on pilot behavior, cognitive engagement, decision-making, and visual scanning during weather avoidance flights. We predict an effect of variations in weather symbology on participant behavior in this study as well, although the direction of effects might be different due to the difference in study designs and participant tasks between this study and the Ahlstrom and Dworsky (2012) study. Furthermore, in the previous study, we did not manipulate individual weather symbols during the simulated flight. In the present study, we examine participant sensitivity to manipulations of individual weather symbols both during a simulated flight and during a change-detection task.

In addition to the broad research goal, we have specific goals that relate to detecting the presence or absence of METAR symbols during a change-detection task (Experiment 2). What we want to estimate is the number of participants (per group) needed to reach a power of 80% to achieve our stated goals.

During our development of the change-detection stimuli (Experiment 2), we used several groups of naïve participants for limited data collection activities (i.e., "pilot" experiments). The purpose of these data collection activities was to fine-tune the display timing for the change-detection task, as well as to gauge the relative response accuracy for the three different WP symbologies. Although each limited data collection activity used a different number of participants, slightly different timing parameters, and different methods (i.e., within- or between-subjects design), we noted the same trend in participant responses, such that the correct detection of METAR symbols was most accurate for WP 3; followed by WP 2; and finally WP 1, which yielded the lowest

detection accuracy. We also noted that the data dispersion (i.e., *SD*) was lowest for WP 3; followed by WP 2; and finally by WP 1, which had the largest dispersion.

Ideally, researchers should plan and conduct their power analyses based on previous research results or from predictions generated by theory. For this study, we have neither of those two alternatives as there are no previous empirical studies and no theory that specifies visual sensitivity to METAR symbols. Alternatively, researchers can generate a data set that exactly follows their research hypothesis—that is, the data reflects what we would expect to find if the hypothetical data were an exact description of real world effects. Figure 9 shows our large hypothetical data set for the detection of METAR symbols based on 500 data points per WP. The relative detection accuracies with the distribution of correct responses (out of six trials) are an extrapolation from the results of our pilot data collection. This is our best estimate of the effects, and we will use this data set to specify three specific study goals that relate to the detection of METAR symbols. In this study, we are mainly interested in comparing the mean detection accuracy among pilots, using the three different WPs, denoted by their group mean μ.
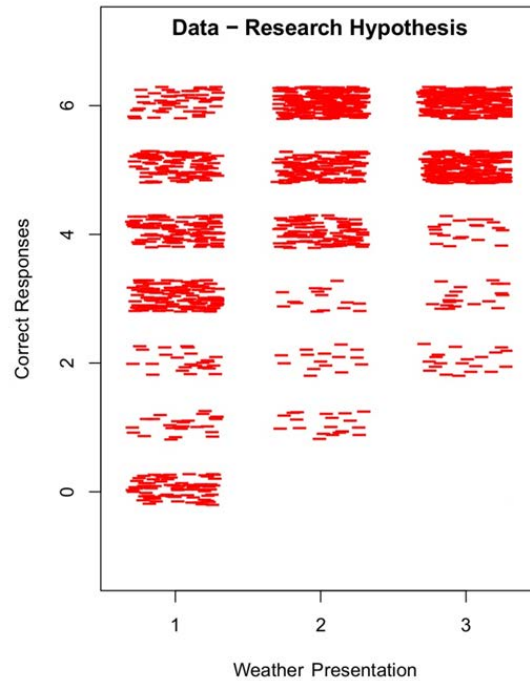
Figure 9. Data generated from a METAR research hypothesis.

For Goal 1 (based on pilot studies), the mean of the WP 2 group exceeds the mean of the WP 1 group, with the 95% HDI excluding the value 0 (i.e., $\mu1 - \mu2 > 0.0$). For Goal 2, the mean of the WP 3 group exceeds the mean of the WP 1 group, with the 95% HDI excluding the value 0 (i.e., $\mu3 - \mu2 > 0.0$). Our third goal relates to the specific symbols used to represent the MEARs in WPs 1-3. WP 1 is using triangles while WP 2 and WP 3 use circles. For Goal 3, the mean of the two groups using circles exceeds the mean of the group using triangles, with the 95% HDI excluding the value 0 (i.e., $\mu1 - (\mu2 + \mu3) / 2 > 0.0$).

Using noncommittal priors, we conducted a Bayesian analysis on the hypothetical data in Figure 9. We repeated this analysis process 400 times—which is equivalent to simulating the METAR experiment 400 times. For each simulated experimental run, we checked the posterior distribution to assess if we achieved our three goals. Our power is the proportion of times we achieve each goal across the 400 repetitions of the hypothetical experiment. The outcome revealed that we would have 80% power in achieving our three goals using 20 participants per group.

2.3 Results and Conclusions

In the following sections, we present results from the cockpit simulation. First, we present an analysis of altitude and heading data recorded during simulation flights. Second, we present an analysis of pilot/controller communication. Subsequently, we present analyses of pilot weather, situation awareness, decision-making, and WP zoom usage. We conclude the section by presenting an analysis of pilot cognitive engagement.

2.3.1 Altitude and Heading Changes

In the following analysis, we assess if there are any credible differences among the three WP groups, regarding flying behavior as measured by altitude and heading changes.

Pilots started each scenario flight at the same altitude, using VFR, but could adjust their altitude based on personal preference or from viewing the OTW view. We used a constant three mile visibility setting. During the VFR flight, most pilots chose to descend to a lower altitude for improved visibility. In addition to these VFR altitude changes, pilots who filed an IFR flight plan were given an IFR altitude by ATC.

Most pilots stayed on the pre-planned route flying from VOR to VOR. However, all pilots made frequent heading changes to *pan* the view on the cockpit weather presentation. Other heading changes were caused by pilot decisions to deviate from the pre-planned route and fly to alternate airport destinations.

The altitude and heading data were sampled at 1 Hz and with 2100 recorded altitude and heading values per pilot (1 Hz x 35 min flight = 2100). For our BANOVA analyses, we used an average altitude and heading value for each pilot.

Figure 10 shows the altitude and heading data for each WP, and Figure 11 shows the posterior contrasts. None of the contrasts for either altitude or heading are credibly different; all posterior distributions have the value 0 included in the 95% HDI.
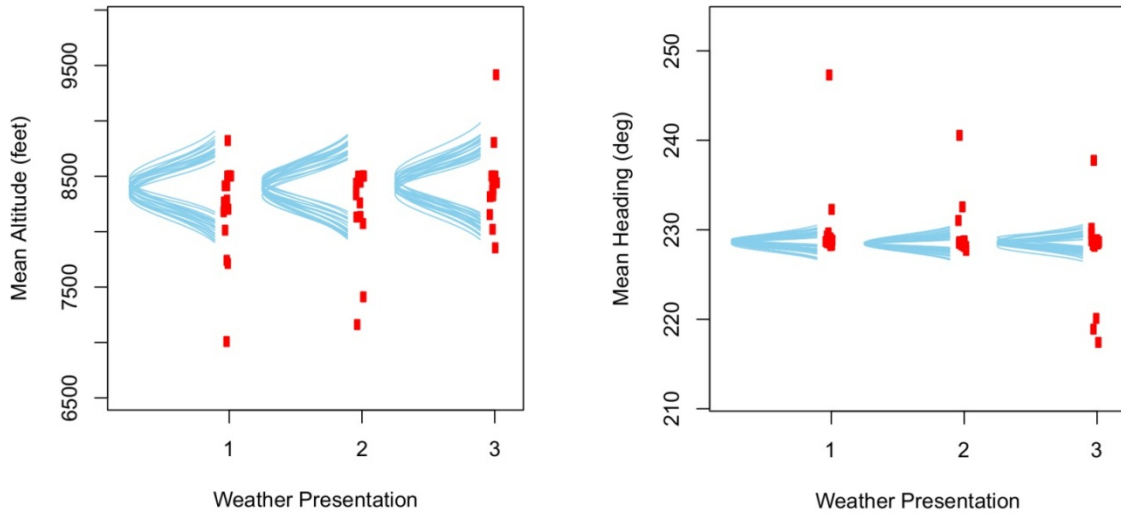
Figure 10. Mean altitude data in feet (left) and mean heading data in degrees (right) for the three WPs.
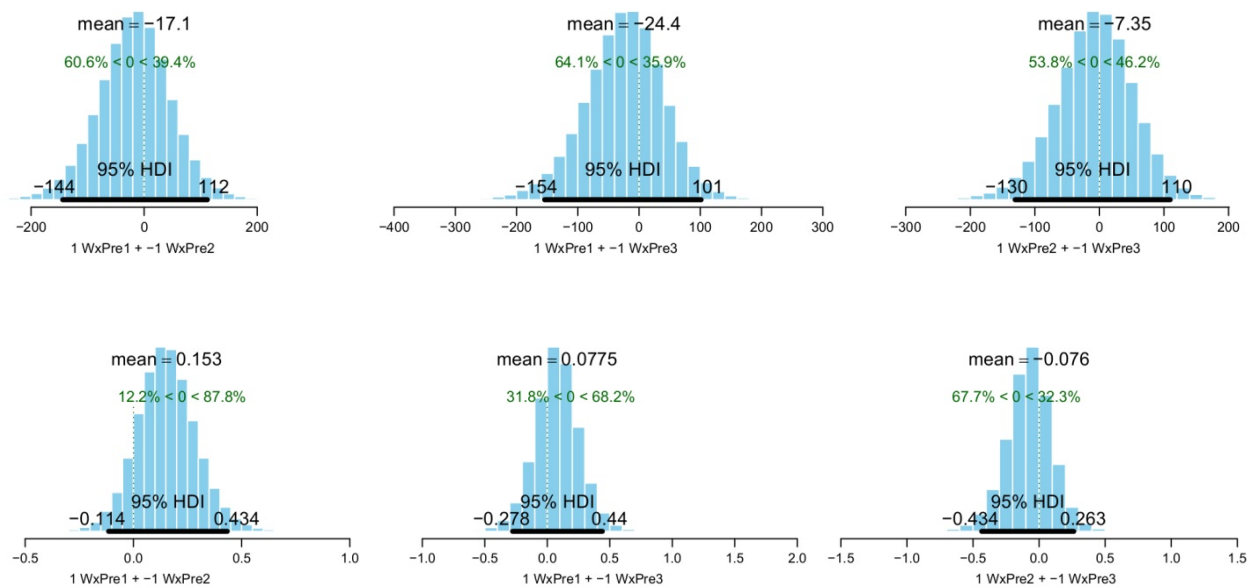
Figure 11. Posterior altitude (top) and heading (bottom) contrasts for the three WPs: WP 1 versus WP 2 (left), WP 1 versus WP 3 (middle), and WP 2 versus WP 3 (right). The black horizontal bar represents the 95% HDI. The vertical dotted axis at 0.00 shows the proportion of the posterior distribution that is below and above the value 0.

To sum up, there are no credible differences in pilot flying behavior between WPs as measured by altitude and heading changes

2.3.2 ATC Communications

During the simulation, we recorded all PTT conversations between the pilots and ATC. Table 7 shows a frequency count of the PTT communications for each WP; Figure 12 shows the associated communication durations. The number of PTT communications was very similar for the three WPs.

Table 7. Frequency Count of PTT Communications per WP

WP	PTT
1	1142
2	1098
3	1103
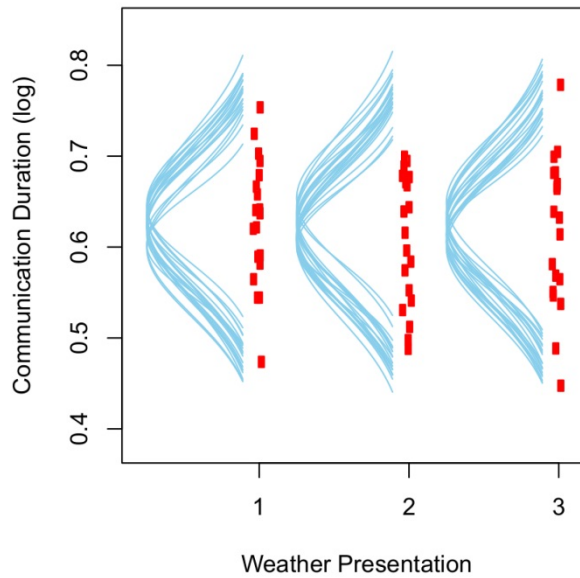
Figure 12. The data (log) and posterior predictive check for WPs 1-3 communication durations.

To analyze the PTT durations, we used all the recorded communications for each pilot and subjected the data to a one-way BANOVA. Figure 13 shows the posterior contrasts for the WP comparison. There are no credible differences in the communication duration between WPs; all posterior distributions include the value 0 within the 95% HDI.
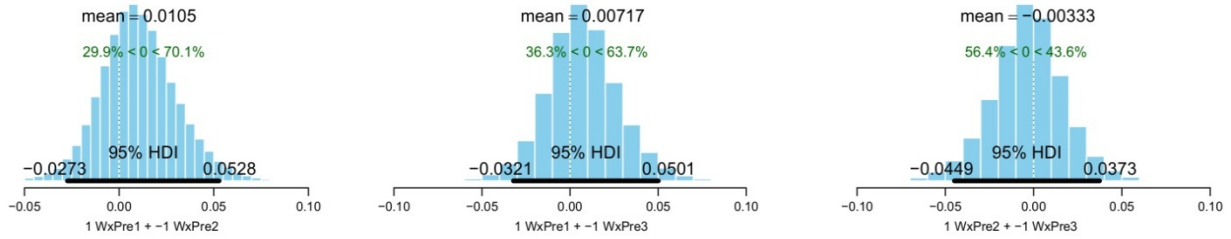
Figure 13. Posterior contrasts for communication durations: WP 1 versus WP 2 (left), WP 1 versus WP 3 (middle), and WP 2 versus WP 3 (right).

To sum up, there are no credible differences in the communication behavior between the three WPs. All pilots exhibited similar communication behaviors. This result is similar to the outcome found by Ahlstrom and Dworsky (2012) for the use of WPs during GA weather avoidance operations.

2.3.3 Weather Situation Awareness - SAGAT Simulation Stops

For the simulation flight, we were primarily interested in pilots' ability to detect the METAR changes introduced at the 10-, 19-, and 30-minute marks. We froze the simulation at the 11-, 20-, and 35-minute marks (SAGAT stops) and assessed whether the pilot detected the METAR change. Figure 14 shows the number of METAR detections for each SAGAT stop and WP.
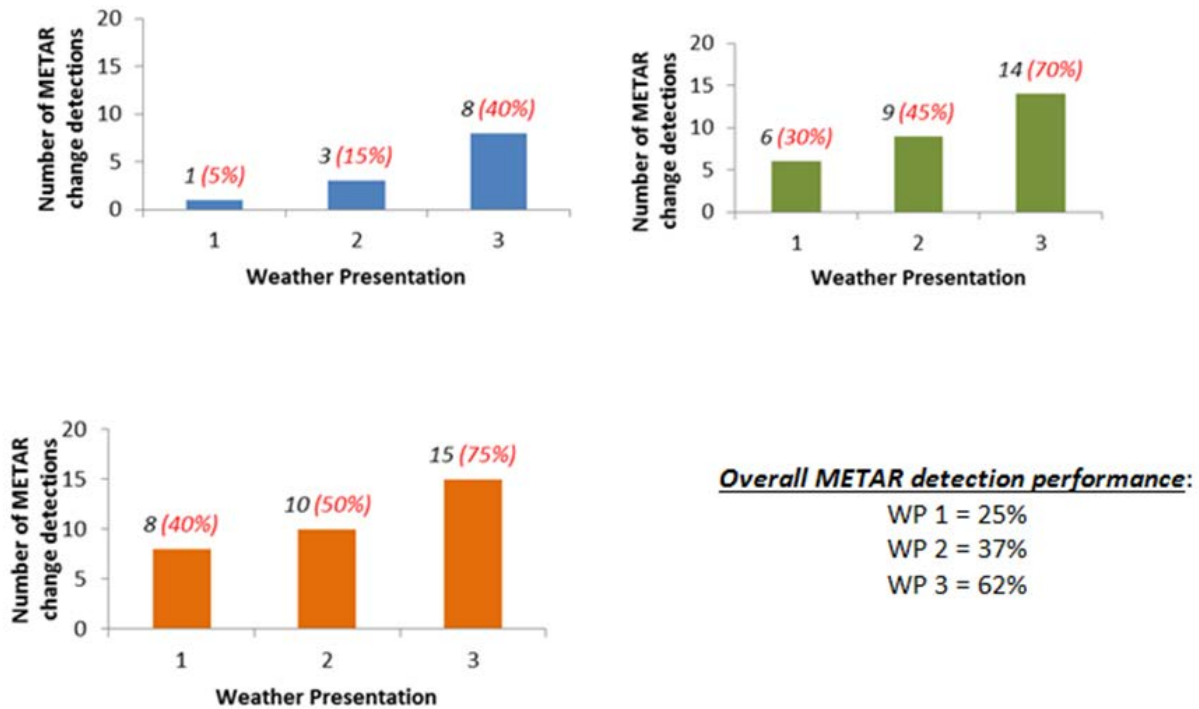
Figure 14. METAR detection data for each WP at the three SAGAT simulation stops. For each WP and SAGAT stop, the maximum number of METAR change detections was twenty (i.e., 20 pilots per group). Top left – the number of METAR change detections (VFR to IFR) at the first SAGAT stop. Top right - the number of METAR change detections at the second SAGAT stop. Bottom left - the number of METAR change detections at the third SAGAT stop. Bottom right – the overall detection performance (%) for each WP (based on 60 opportunities per WP).

29

As can be seen in Figure 14, the METAR change-detection was generally poor. The overall detection performance for pilots using WP 1, WP 2, and WP 3 was 25%, 37%, and 62%, respectively.

From the METAR detections, we computed an overall detection score for each pilot that is based on the number of detections across the three METAR changes. Figure 15 (left) shows the detection data for each of the three WPs, with the detection accuracy being a function of whether the pilot detected zero, one, two, or all three of the METAR changes. Figure 15 (right) also shows the posterior distribution after the Bayesian analysis. As we can see, the mean detection performance is highest for WP 3; followed by WP 2; and finally WP 1, which yields the lowest detection accuracy.
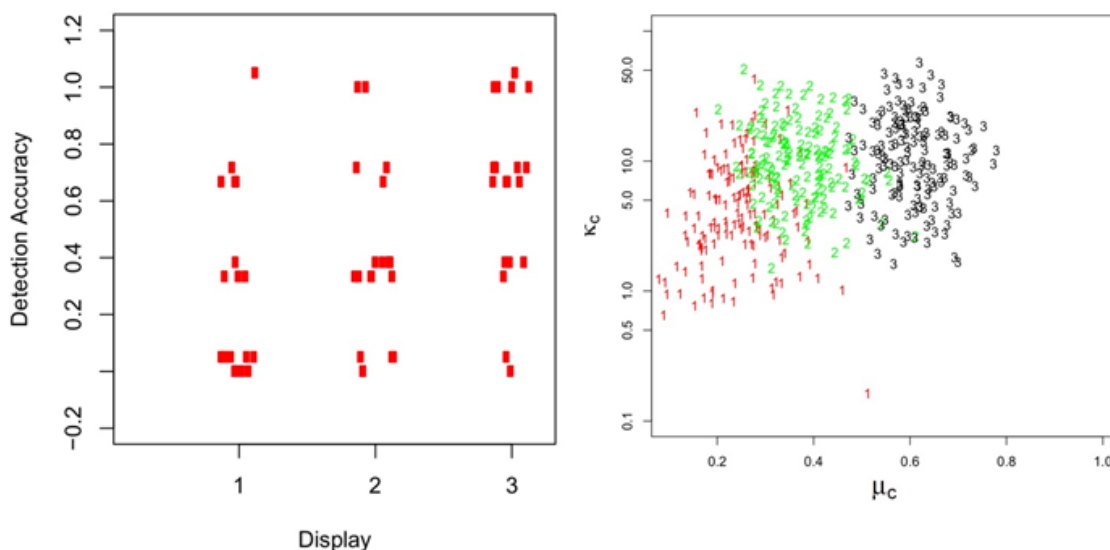
Figure 15. Left - detection accuracy data from the simulation flight for each of the three WPs (1-3).

Figure 16 shows histograms of the posterior contrasts with the difference between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right). Because the value 0 is included in the 95% HDI for the contrast between WP 1 and 2 WP, these two WPs are not credibly different. For the contrast between WP 1 and WP 3, however, we have a credible difference with a higher detection performance for WP 3 compared to WP 1. We also have a credible difference between WP 2 and WP 3, with WP 3 having a higher detection accuracy compared to WP 2.
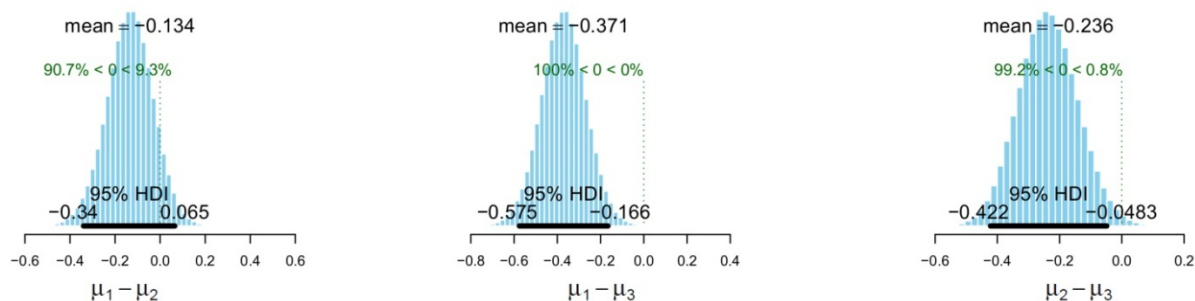
Figure 16. Posterior contrasts for the difference in METAR detections between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

We also analyzed pilot detection performance in terms of each WP symbol and color combination. The METAR symbol for WP 1 is a triangle, and for WP 2 and WP 3 it is a circle. The METAR-symbol color change from VFR to IFR for WP 1 and WP 2 is blue to yellow, and for WP 3 it is white to red. Figure 17 shows the posterior distributions for the symbol and color contrasts. There is a credible difference in the detection accuracy between triangles and circles with circles, on average, yielding higher detection performance than triangles. There is also a credible difference in the detection performance between the blue/yellow and the white/red color change, with the white/red color change, on average, yielding a higher detection performance than the blue/yellow color change.
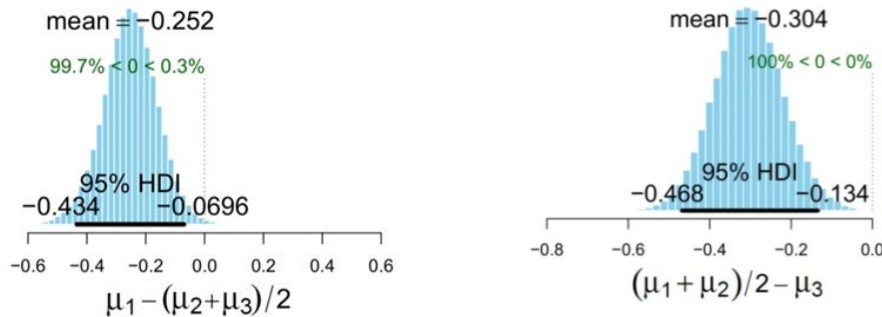
Figure 17. Posterior contrasts for the detection of METAR symbols defined by triangles versus METAR symbols defined by circles (left), and the difference in detection between blue/yellow and white/red METAR symbols (right).

A factor that could affect METAR detection performance (as illustrated in Figure 15) is experience with cockpit weather symbology. Theoretically, pilots who currently use or have experience with electronic presentations of weather symbols could have a higher propensity to detect symbol changes. Conversely, pilots who lack this experience could have a lower propensity to detect changes in a symbol's status (color or shape). To address this issue, we used data from the Biographical Questionnaire to assess each pilot's experience with electronic weather symbols (e.g., ADS-B, Garmin, ForeFlight, XM weather, and so forth) to see if it correlated with the METAR detection accuracy. Surprisingly, there were only 17 pilots who reported experience with electronic weather symbol presentations. All other pilots reported having no personal experience with electronic weather symbols. Nine of the pilots with prior experience were using WP 1, with four pilots with prior experience using WP 2 and WP 3, respectively.

When analyzing the METAR detection performance for SAGAT stop 1 (as shown in Figure 14, top left), we find that the single pilot for WP 1 who detected the METAR change did have prior experience using weather symbology. The three pilots who detected the METAR change using WP 2, however, had no prior experience in the use of electronic weather symbols. For WP 3, only two of the eight pilots who detected the METAR change had prior experience. Therefore, the detection performance for METAR change 1 cannot be explained in terms of pilot experience with electronic weather symbols.

Analyzing pilot experience and detection performance for SAGAT stop 2 (as shown in Figure 14, top right) revealed that four of the six pilots who detected the METAR change using WP 1 had prior experience. For WP 2, only one of the nine pilots who detected the METAR change had prior experience. For WP 3, four of the fourteen pilots who detected the METAR change had prior experience. Again, prior experience with electronic weather symbols does not seem to account for the METAR detection performance.

For SAGAT stop 3 (Figure 14, bottom left), we find that of all the pilots who detected the METAR change, WP 1 and WP 2 had three experienced pilots each, but there were four experienced pilots using WP 3. Again, the detection performance for METAR change 3 cannot solely be explained by pilot experience with electronic weather symbols.

Of all the METAR change-detections in Figure 14, there were only eight pilots who detected all three METAR changes during the flight. One of these pilots was using WP 1; two were using WP 2, with the remaining five pilots using WP 3. All pilots for WP 1 and WP 2 had prior experience with electronic weather symbols. For WP 3, however, only two of the five pilots had prior experience.

To sum up, there are credible differences in the METAR detection accuracy between pilot groups using different weather presentations. Although there is modest overall detection performance for pilots using WP 3, the detection performance was poor, at best, to METAR changes for pilots using WP 1 and WP 2. With regards to METAR symbology, METAR symbols using circles and a white to red color change (VFR to IFR) yield higher detection performance than METAR triangle or circle symbols with a blue to yellow color change. Prior experience with modern electronic weather symbols cannot account for the METAR detection performance.

2.3.4 Decision Making - Weather, Deviation, and IFR Requests

When a pilot detected a METAR change (from VFR to IFR) during flight, it could affect the pilot's decision-making in a number of ways. For example, the pilot could decide to continue the flight using VFR, the pilot could contact ATC and request an IFR flight plan, or the pilot could contact ATC and inquire about weather updates for the destination airport (or alternate airports).

One question of interest is whether pilots differ systematically in their decision-making based on their specific WP. Table 8 presents the number of pilot requests for weather information (Weather), deviations to alternative airports (Deviation), and for requesting IFR flight plans (IFR). As is show in Table 8, there were very few requests per WP overall with small numerical differences between WPs. Although there was no credible main effect of WP, WP 2 and WP 3 have a larger total number of requests compared to WP 1.

Table 8. Frequency Count of Weather, Deviation, and IFR Requests per WP

WP	Weather	Deviation	IFR
1	10	3	8
2	14	5	11
3	18	6	9

Comparing the frequency counts across the three request types, both WP 2 (posterior mean = 0.273, 95% HDI from 0.018 to 0.522) and WP 3 (posterior mean = 0.333, 95% HDI from 0.080 to 0.581) have credibly more weather requests than deviation requests.

We also wanted to see if the detection of at least one of the three METAR changes affected a pilot's propensity to make ATC requests. Table 9 shows the frequency count for each WP in Table 8 in terms of whether requests came from pilots who detected versus pilots who did not detect at least one of the three METAR changes. Only pilots with no METAR detections from WP 1 and WP 2 made requests, with only one request for WP 2 and seven requests for WP 1 (see Table 9). Of the seven requests made for WP 1, five were IFR requests along with weather and a deviation request. For WP 2, the single request was a weather request. There is a credible difference in the frequency counts for pilots who detected METARs between WP 1 and WP 2 (mean posterior = -0.284; 95% HDI from -0.50 to -0.08), with a higher count for WP 2, and between WP 1 and WP 3 (mean posterior = -0.32; 95% HDI from -0.52 to -0.13), with a higher count for WP 3.

Table 9. WP Frequency Count of the Total Number of Requests for Pilots who Detected/did not Detect at Least one METAR Change

WP	METAR Detection	No METAR Detection
1	14	7
2	29	1
3	33	0

In addition to the similarity across WPs for the number of requests, the points in the scenario at which the requests occurred were also very similar. Weather requests for all three WPs, on average, occurred 18-20 minutes into the scenario. IFR requests, on average, occurred 19-21 minutes into the scenario and deviation requests, on average, occurred at 24-28 minutes into the scenario.

2.3.5 Weather Presentation Usage - Zoom Changes and Zoom Durations

In this section, we assess how pilots used the weather presentation zoom by analyzing zoom display durations and zoom level transitions (going from one zoom level to another). We use a two-way BANOVA to analyze the zoom display durations with factors WP and zoom level. For this analysis, we used all the recorded zoom level durations for each pilot (see Figure 18).
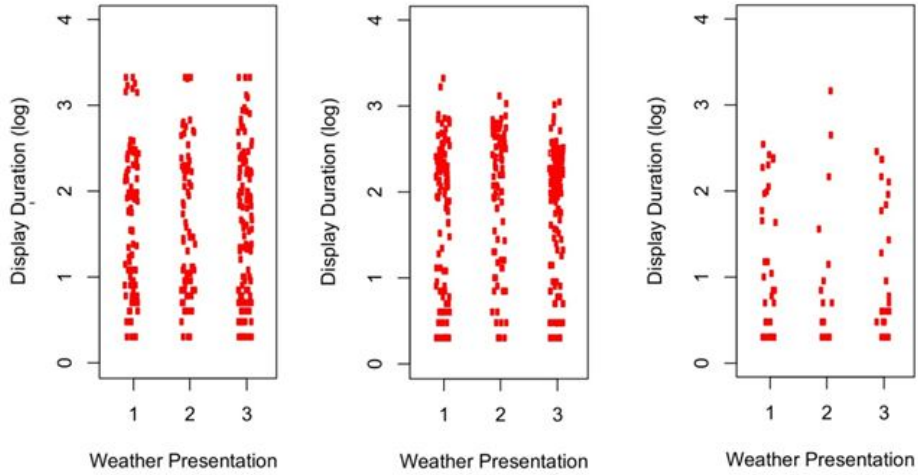
Figure 18. WP display durations (log) data for zoom level 1 (left), zoom level 2 (middle), and zoom level 3 (right).

Figure 19 shows the posterior distributions for zoom duration contrasts between WP (top) and zoom levels (bottom). There were no credible differences in zoom durations between WPs; all three contrasts include the value 0 within the 95% HDI. For zoom levels, there is a credible difference between the display durations for zoom level 1 (5 nmi range rings) and zoom level 2 (20 nmi range rings) with zoom level 2 being displayed for longer durations than zoom level 1 (bottom left). There is also a credible difference between the display durations for zoom level 1 and zoom level 3 (50 nmi range rings) with zoom level 1 being displayed for longer durations than zoom level 3 (middle). There is also a credible difference between the display durations for zoom level 2 (20 nmi range rings) and zoom level 3 with zoom level 2 being displayed for longer durations than zoom level 3 (bottom right). Finally, there were no credible differences between the interaction of WP and zoom levels.
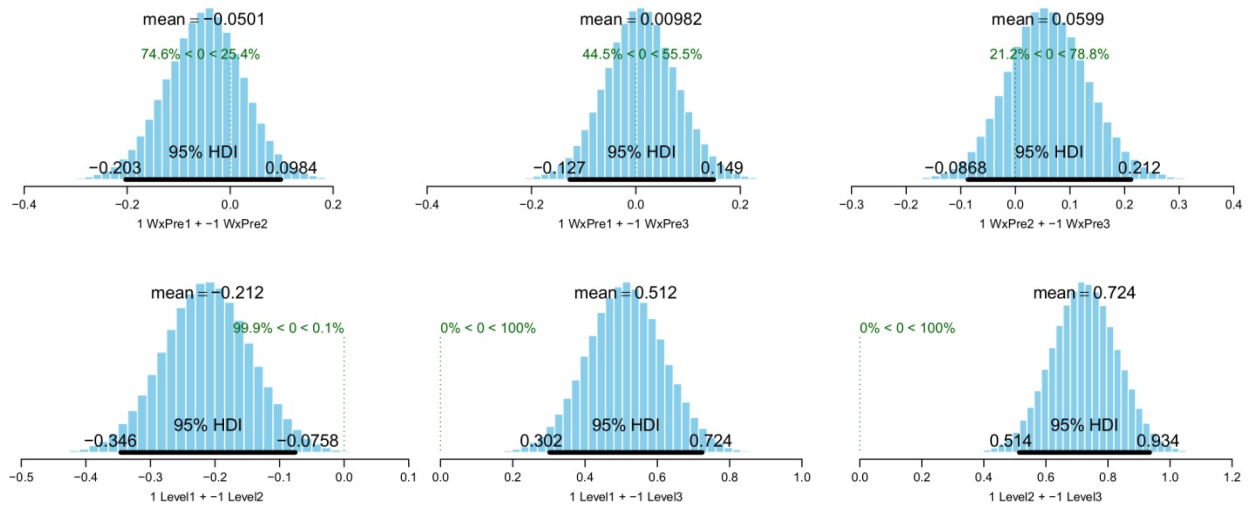
Figure 19. Posterior contrasts for differences between WPs (top) and zoom levels (bottom) on log zoom durations.

34

Table 10 shows the frequency count data for zoom level transitions and WP. In Table X, "1-2" denotes a transition from zoom level 1 to zoom level 2, "2-3" denotes a transition from zoom level 2 to zoom level 3, "3-2" a transition from zoom level 3 to zoom level 2, and "2-1" denotes a transition from zoom level 2 to zoom level 1.

Table 10. Frequency Count of Zoom Level Transitions per WP

WP	Transition between zoom levels			
	1→2	2→3	3→2	2→1
1	89	51	48	75
2	70	31	31	59
3	109	41	40	96

The zoom level transition counts are very similar across WPs and transition levels with no credible differences between WPs.

The result of this weather presentation analysis is similar to the result found by Ahlstrom and Dworsky (2012) for GA weather avoidance operations. They found that different pilot groups in the study exhibited the same zoom display behavior and they also displayed each zoom level for similar durations.

2.3.6 Cognitive Engagement

Ahlstrom and Dworsky (2012) found that GA pilot fNIR oxygenation levels were higher during IFR portions of flights than during VFR portions of flights. This higher cognitive engagement during IFR flights can be attributed to the difference between VFR and IFR pilots with regards to the use of instruments, flight planning, ATC communication, and flight procedures. For the present analysis, we are interested in assessing the effect of WP on pilot oxygenation levels. We also want to know whether pilots who detected METAR changes are more cognitively engaged in planning and decision-making (as indicated by increased oxygenation levels) compared to pilots who did not detect METAR changes. First, we analyze the oxygenation levels for pilots who detected METAR changes. Next, we assess the oxygenation levels for pilots who did not detect METAR changes. Finally, we compare the oxygenation levels for pilots who detected versus pilots who did not detect the METAR changes.

For the initial analysis, we are mainly interested in the relative oxygenation levels before and after a METAR change and we are only using the oxygenation data for pilots who detected METAR changes. First, for each pilot we averaged the oxygenation values across the 16 fNIR channels, aiming for an overall prefrontal oxygenation assessment rather than looking at specific prefrontal regions or differences between the left and right hemisphere. Second, we used all the recorded data (2 Hz) 1 min before and 1 min after the METAR change. For each before and after value, we calculated a difference score that was used for the BANOVA analyses. Therefore, for each successful METAR change-detection by a pilot, we used 120 oxygenation values in the analysis.

2.3.6.1 Oxygenation levels for pilots who detected METAR symbol changes – Effects of WP and METAR change.

Figure *20* shows the oxygenation data for the three METAR changes and the three WPs. Using WP (1-3) and METAR change (1-3) as factors in a two-way BANOVA, we computed differences in oxygenation between the three WPs at the three METAR changes.
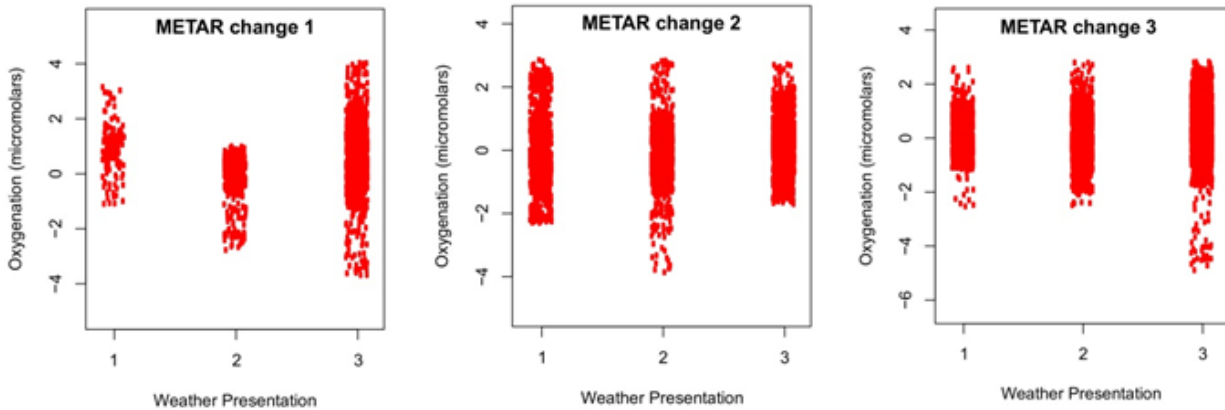
Figure 20. Oxygenation data for the three WPs and the three METAR changes. The oxygenation data are for pilots who detected the three METAR changes.

Figure 21 shows the posterior distributions for the main effect of WP. There is a credible difference between WP 1 and WP 2, with WP 1 having a higher oxygenation level than WP 2. There is no credible difference between WP 1 and WP 3. However, there is a credible difference between WP 2 and WP 3 with WP 3 having a higher oxygenation level than WP 2.
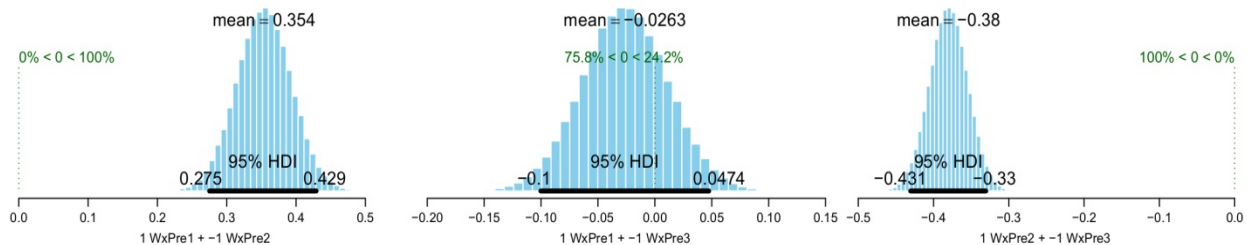
Figure 21. Posterior contrasts for the main effect of WP. Left, the comparison between WP 1 and WP 2. Middle, the comparison between WP 1 and WP 3. Right, the comparison between WP 2 and WP 3.

Figure 22 shows the contrasts for the main effect of METAR change. All three METAR changes produced different levels of oxygenation. First, there is a credible difference between METAR change 1 and 2, with a higher oxygenation for METAR change 1 compared to METAR change 2. Second, there is a credible difference in oxygenation between METAR change 1 and 3, with METAR change 1 having a higher oxygenation than METAR change 3. Third, there is a credible difference between METAR change 2 and 3, with METAR change 3 having a higher oxygenation than METAR change 2.
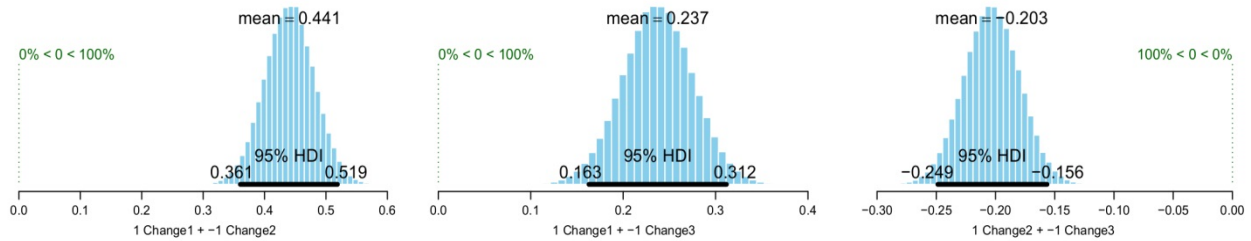
Figure 22. Posterior contrasts for the main effect of METAR change on oxygenation. The left histogram shows the difference in oxygenation between METAR change 1 and METAR change 2. The middle histogram shows the difference between METAR change 1 and METAR change 3. The right histogram shows the difference between METAR change 2 and METAR change 3.

Contrasting the three WPs and the three METAR changes we only find credible differences in oxygenation for METAR change 1 (i.e., 10 minutes into the scenario). Figure 23 shows the posterior distributions for contrasts between the three WPs and METAR change 1. There is a credible difference between WP 1 and WP 2, with WP 1 having a higher oxygenation level than WP 2. There is no credible difference between WP 1 and WP 3, but a credible difference between WP 2 and 3 with WP 3 having a higher oxygenation level than WP 2. There were no credible differences between WPs for METAR change 2 and 3; all posterior distributions included the value 0 within the 95% HDI.
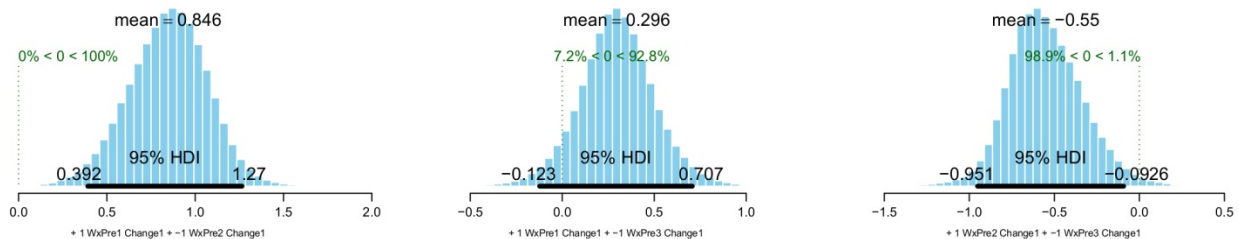
Figure 23. Posterior contrasts for the effect of WP and METAR change 1. The left histogram shows the difference between WP 1 and WP 2, the middle histogram shows the difference between WP 1 and WP 3, the right histogram shows the difference between WP 2 and WP 3. METAR change 1 (VFR to IFR symbol change at the destination airport) occurred 10 minutes into the scenario flight.

2.3.6.2 Oxygenation levels for pilots who did not detect METAR symbol changes – effects of WP and METAR change.

We also analyzed the oxygenation data for the pilots who did not detect any of the three METAR changes. For these analyses, we are not interested in the METAR change per se (because pilots did not see it), but instead we assess the relative oxygenation for these pilots during the time period before and after each change.

Figure 24 shows the oxygenation data for pilots who did not detect the METAR changes. Using WP (1-3) and METAR change (1-3) as factors, and using all the recorded data 1 min before and 1 min after the METAR change, we subjected the difference scores to a two-way BANOVA. Figure 25 shows the posterior distributions for the main effect of WP. None of the three contrasts are credibly different; all 95% HDIs include the value 0.
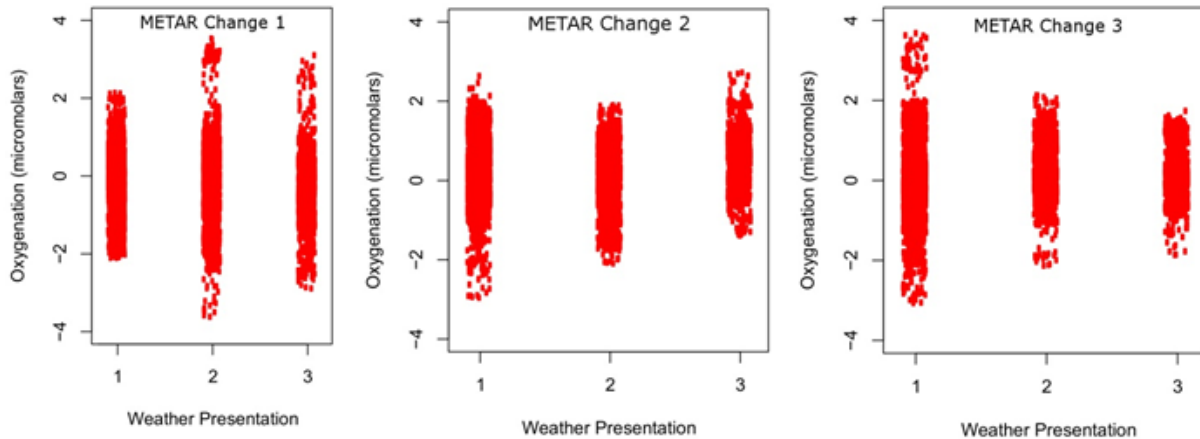
Figure 24. Oxygenation data for pilots who did not detect the METAR changes, by WPs and METAR change.
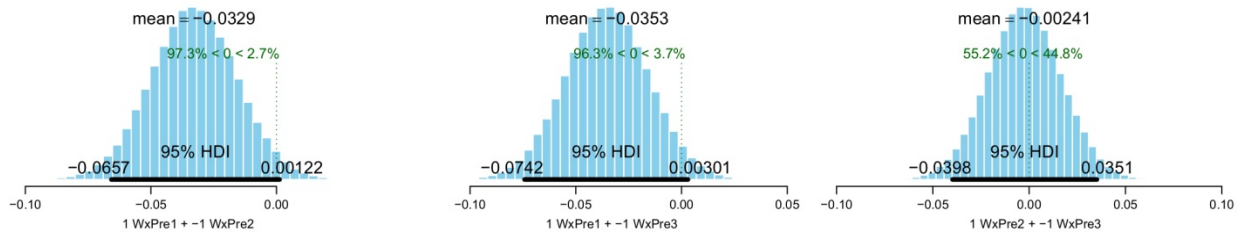
Figure 25. Posterior contrasts for the main effect of WP for pilots who did not detect METAR changes. Left, the comparison between WP 1 and WP 2. Middle, the comparison between WP 1 and WP 3. Right, the comparison between WP 2 and WP 3.

Figure 26 shows the posterior distributions for contrasts of the main effect of METAR change time (one minute before and one minute after the change) on pilots who did not detect the METAR status changes. There is a credible difference between the time durations for METAR change 1 and METAR change 2 on oxygenation, with a higher oxygenation level for the time period at METAR change 2. There is also a credible difference between the time durations for METAR change 1 and METAR change 3, with a higher oxygenation level for the time period at METAR change 3. Finally, there is a credible difference between the time durations for METAR change 2 and METAR change 3, with a higher oxygenation level for the time period at METAR change 2.
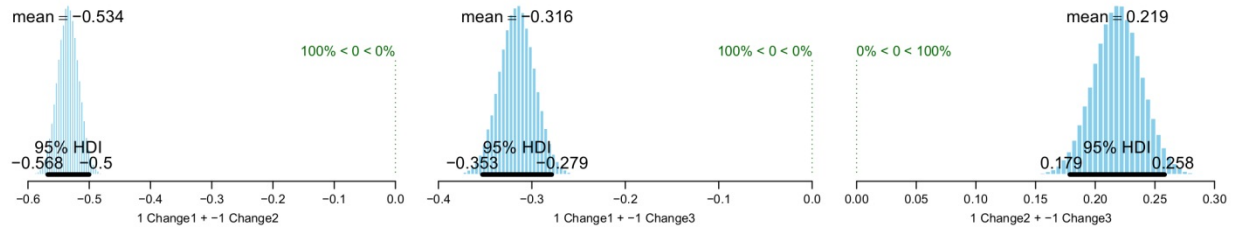
Figure 26. Posterior contrasts for the main effect of METAR change time on oxygenation for pilots who did not detect the METAR changes. The left histogram shows the difference in oxygenation between METAR change 1 and METAR change 2. The middle histogram shows the difference between METAR change 1 and METAR change 3. The right histogram shows the difference between METAR change 2 and METAR change 3.

Besides assessing the main effects of WP and METAR change times, we also contrasted the three WPs with the three METAR change times. Although there are no credible differences among WPs for the METAR change 3 time duration, there are credible differences in oxygenation for the METAR change 1 and the METAR change 2 time durations.

Figure 27 shows the posterior contrasts for METAR change 1 (top) and METAR change 2 (bottom). For METAR change 1 there is a credible difference between WP 1 and WP 3 with WP 1 having a higher oxygenation than WP 3. There is also a credible difference between WP 2 and WP 3 with WP 2 having a higher oxygenation than WP 3. For METAR change 2 there is a credible difference between WP 1 and WP 3 with WP 3 having a higher oxygenation than WP 1. There is also a credible difference between WP 2 and WP 3 with WP 3 having a higher oxygenation than WP 2.

Figure 27. Posterior contrasts for the interaction effects of WP and METAR change 1 (top) and WP and METAR change 2 (bottom) for pilots who did not detect the METAR changes. The left side histograms show the difference between WP 1 and WP 2, the middle histograms show the difference between WP 1 and WP 3, the right side histograms show the difference between WP 2 and WP 3.

2.3.6.3 Comparing oxygenation levels at METAR change times for pilots who detected versus pilots who did not detect symbol changes.

Pilots perform multiple tasks and this yields a relative cognitive engagement during different phases of flight. Even for pilots that did not detect the METAR changes these pilots were still piloting (i.e., navigating, planning) and therefore actively involved in decision-making while in flight. What is surprising, however, is the drastic difference in oxygenation during the METAR change periods between the pilots who detected the changes versus the pilots who did not detect the changes. If we compare the contrasts in Figure 22 (pilots who detected changes) and Figure 26 (pilots who did not detect changes), we see that the contrast effects are in the opposite direction. In Figure 22, METAR change 1 yields a higher oxygenation level than change 2. For the pilots who did not detect the METAR changes, the change 2 time period yields a higher oxygenation level than the change 1 period. Figure 22 also shows that METAR change 1 yields higher oxygenation than change 3, while for the pilots who did not detect the changes the METAR change 3 period yields higher oxygenation than change 1. The reversed order is also true for the contrast between change 2 and change 3. Although Figure 22 shows that change 3 yields a higher oxygenation than change 2, for the pilots who did not detect the METAR changes, the effect is in the opposite direction. The change 2 time period yields a higher oxygenation than the change 3 time period. Clearly, cognitive engagement differs between pilots who detected METAR changes and pilots who did not detect METAR changes, as demonstrated by the credible differences in pre-frontal oxygenation levels.

To analyze this further, we used the factors METAR detection (no detection versus detection) and METAR change (1-3) in a two-way BANOVA to contrast pilots who detected METAR changes to assess if these pilots had an increased oxygenation level compared to pilots who did not detect METAR changes. When pilots detected a METAR status change (i.e., VFR to IFR) during flight, the METAR change informed pilots about a reduction in an airport's ceiling and visibility conditions. Therefore, this information could potentially trigger pilot decisions related to requesting additional weather information from ATC, decisions about continuing the flight VFR versus IFR, continuing towards the destination airport, selecting a new destination airport, or whether to contact ATC and request an IFR flight plan.
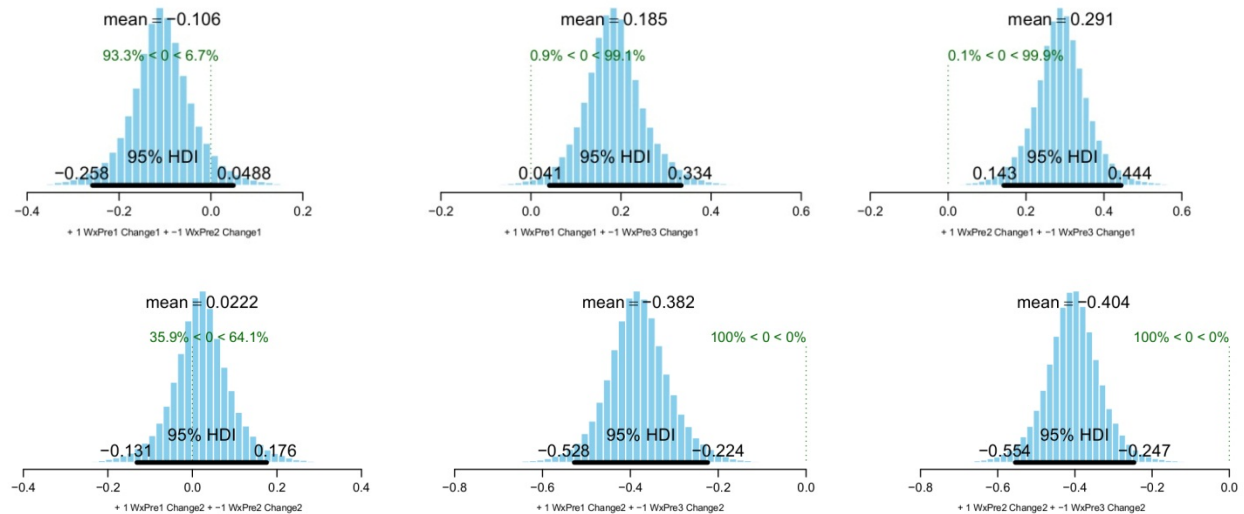
Figure 28 shows the posterior contrasts for the main effect of METAR change on oxygenation levels for pilots who detected the METAR changes versus pilots who did not detect the METAR changes. There is a credible difference between pilots who detected versus pilots who did not detect METAR Change 1, with a higher oxygenation level for pilots who detected the change compared to pilots who did not detect the change. Because METAR Change 1 involved the pre-planned destination airport (KMRB), pilots who detected the change at 9 minutes into the flight were more likely to engage in decision-making and planning regarding their continuing flight (e.g., ATC weather requests, VFR versus IFR, alternate airports) compared to the pilots who did not detect the change. We would expect this additional decision-making and planning to be reflected by heightened cognitive engagement, as measured by oxygenation levels in the prefrontal cortex.
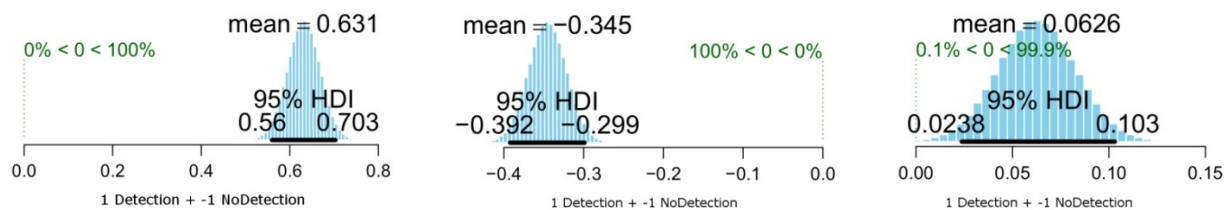
Figure 28. Posterior contrasts for the main effect of METAR change 1 (left), METAR change 2 (middle), and METAR change 3 (right) on oxygenation levels for pilots who detected the change (Detection) versus pilots who did not detect the change (NoDetection).

For METAR change 2 we have the opposite effect, with a higher oxygenation level for pilots who did not detect the METAR change compared to pilots who detected the METAR change. Many of the pilots who detected METAR change 1 had already picked HGR (Hagerstown) as a potential alternative airport and had already made some decisions about their continuing flight prior to METAR change 2. Therefore, because METAR change 2 did not involve HGR it likely did not add additional cognitive load on these pilots. For the remainder of the pilots who did not detect METAR change 2 it is not obvious why this group of pilots had a higher oxygenation level compared to the pilots who detected the change. One possible reason, due to the fact that these pilots did not detect METAR change 1, is that this group of pilots did not make many flight decisions prior to METAR change 2 but at 19 minutes into the scenario they were getting closer to the destination airport and were therefore more cognitively engaged in flight planning.

Finally, there is a credible effect of METAR change 3 on oxygenation with a higher oxygenation level for pilots who detected the change versus pilots who did not detect the change. METAR change 3 involved HGR (Hagerstown) which was chosen by many pilots as an alternative airport already at METAR change 1. When HGR's METAR symbol indicated IFR, new decisions had to be made including asking ATC about the current surface weather at HGR and other nearby airports, and selecting and reviewing relevant approach plates.

2.4 Discussion

There are no credible differences in pilot flying behavior between the three WPs as measured by altitude and heading changes, and all pilots exhibit a similar communication behavior. There is also a similarity across WPs for the number of weather, deviation, and IFR requests and the points in the scenario at which the requests occur. Pilot's use of the WP zoom functionality (i.e., zoom level transition counts) is also very similar with no credible differences between WPs.

However, there are credible differences in the METAR detection accuracy between pilot groups using WPs. Although there is modest overall detection performance for pilots using WP 3, the detection performance was poor, at best, to METAR changes for pilots using WP 1 and WP 2. METAR circle symbols with a white to red color change (VFR to IFR) yield higher detection performance than METAR triangle or circle symbols with a blue to yellow color change. Prior experience with modern electronic weather symbols cannot account for this performance.

Pilots who detected symbol changes had a credibly different oxygenation level compared to pilots who did not detect symbol changes. In most cases, detecting a symbol change increased the pre-frontal blood oxygenation—which is symptomatic of an increased cognitive engagement like flight planning and decision-making.

3. EXPERIMENT 2

Experiment 1 addressed, specifically, pilot perception of METAR-symbol color changes during a realistic and representative piloting task. In order to focus on detection of METAR-symbol color changes, we intentionally did not manipulate changes to other weather graphics such as precipitation, SIGMET, and lightning symbols during the simulated flight. There is, however, a need to assess the effect of different WP symbologies on change-detection performance for these other weather graphics. To accomplish this, we conducted a non-simulator experiment (Experiment 2) that examined basic change-detection performance in a more controlled manner. In contrast to Experiment 1—in which each pilot's primary task was flying the plane and, therefore, change-detection was implicit—the primary task in Experiment 2 was detecting changes to a range of WP elements in static weather images (i.e., explicit change-detection).

In addition to the four weather information symbologies (i.e., precipitation, lightning, METAR, and SIGMET) used in Experiment 1, we also included time-stamp information (see Appendix E) in Experiment 2. Although commercial WP symbologies all include similar information for weather information elements (FAA, 2010), there is less of a consensus regarding the format and location for time-stamp information. In this experiment, we are exploring the effect on detectability from one particular time-stamp format and one particular time-stamp location.

3.1 Method

3.1.1 Participants

Sixty instrument-rated (56 male and 4 female) and four non-instrument-rated (all male) GA pilots volunteered to participate in the study. The sixty instrument-rated pilots were those who completed the simulation study (i.e., Experiment 1). The participants were recruited from the pool of federally employed and contract pilots at the FAA WJHTC. Participants were paid at their regular hourly rate while participating. Participants were randomly assigned to one of three WPs (Presentation 1, $n = 21$; Presentation 2, $n = 21$; Presentation 3, $n = 22$).

3.1.2 Testing Facility

The part-task study was conducted in the Cockpit Simulation Facility at the FAA WJHTC. All testing was conducted using a purpose-built computerized testing facility comprising 12 cubicles, each equipped with a desktop computer (Hewlett-Packard HP Pro 3500) and a 22-inch LCD monitor (Dell P2212H) set at a resolution of 1920 × 1080 pixels. To facilitate responses during the experimental task, the "z" and "/" keyboard keys were labeled *Yes* and *No,* respectively.

3.1.3 Materials

The visual stimuli consisted of static WP images (428 × 1021 pixels) that were similar, visually, to the dynamic cockpit WP employed in the simulator study. At a viewing distance of 64 cm, the viewing angle of the WP images subtended 9° horizontally and 20° vertically.

A single, *complete* WP image was used as the basis for all stimuli in this experiment. In addition to the underlying terrain, the complete image contained the following weather-information elements:

- Aviation routine weather report for a specific location (METAR). Small, color-coded symbols were used to summarize METAR as either Visual Flight Rules (VFR) or Instrument Flight Rules (IFR) flight conditions, according to visibility and ceiling. In the complete image, approximately half of the METARS indicated VFR conditions and the other half indicated IFR conditions.

- Significant Meteorological Advisory (SIGMET) information that depicted advisories on weather that is significant to the safety of all aircraft. The region(s) affected by the SIGMET was enclosed by a polygon (e.g., rectangle).

- Lightning strikes. Regions affected by lightning strikes are marked by small symbols.

- Precipitation, which depicts the intensity of precipitation overlaid on the map.

- Time-stamp, which contained a date and time, and the duration (in minutes), since the weather display was last updated. Note that the data contained within the time-stamp were not changed on any of the images.

We used the GNU Image Manipulation Program (www.gimp.org) to create a set of *changed* images by digitally removing, or changing the color of, weather-information elements in the complete image. (For a step-by-step tutorial on how to produce change-detection images with GIMP; see Ball, Elzemann, & Busch, 2013.) Each changed image incorporated a change (i.e., removal/color change) to a single weather-information element only. The set of changed images comprised the

- complete image with the coloring removed from all METARs. A METAR without a fill color indicates that the reporting station is out of order, or is otherwise not currently transmitting routine weather reports.

- complete image with all METARs set to indicate IFR conditions.

- complete image with all METARs set to indicate VFR conditions.

- complete image with the SIGMET removed.

- complete image with all lightning strikes removed.

- complete image with all precipitation removed.

- complete image with the time-stamp removed.

We then created 12 unique "change trials" by pairing specific images. In each change trial, the change was either an: (a) onset (i.e., appearance) of a weather-information element; (b) offset (i.e., disappearance) of a weather-information element; or (c) change in color of the METARs. The change trials—including the type of change and the image pairing used to create the change—are described in Table 11. The actual image pairs are presented in Appendix F and presented along with the data and results in the Results section.

Table 11. List of Change Trials by Weather-Information Element Changed, Type of Change, and Image Pair

Weather-information element changed	Type of change	Image 1	Image 2
METARs	Onset	Complete minus METARs colors	Complete
METARs	Offset	Complete	Complete minus METARs colors
METARs	Color: IFR→VFR	Complete with all METARs = IFR	Complete with all METARs = VFR
METARs	Color: VFR→IFR	Complete with all METARs = VFR	Complete with all METARs = IFR
SIGMET	Onset	Complete minus SIGMET	Complete
SIGMET	Offset	Complete	Complete minus SIGMET
Lightning	Onset	Complete minus lightning	Complete
Lightning	Offset	Complete	Complete minus lightning
Precipitation	Onset	Complete minus precipitation	Complete
Precipitation	Offset	Complete	Complete minus precipitation
Time-stamp	Onset	Complete minus time-stamp	Complete
Time-stamp	Offset	Complete	Complete minus time-stamp

In addition to the 12 change trials, we created eight catch (i.e., no-change) trials (see Table 12 and Appendix E).

Table 12. List of Catch Trials

Catch trial	Image 1	Image 2
1	Complete	Complete
2	Complete minus METARs colors	Complete minus METARs colors
3	Complete with all METARs = IFR	Complete with all METARs = VFR
4	Complete with all METARs = VFR	Complete with all METARs = IFR
5	Complete minus SIGMET	Complete minus SIGMET
6	Complete minus lightning	Complete minus lightning
7	Complete minus precipitation	Complete minus precipitation
8	Complete minus time-stamp	Complete minus time-stamp

Finally, we created a set of practice images using the underlying terrain from the complete image. Instead of weather information, we used colored generic shapes (i.e., squares, rectangles, stars, blobs) to create eight change (i.e., onset and offset) and six catch trials (see Appendix E).

3.1.4 Independent, Between-Subjects Variable: Weather Presentation (WP)

We manipulated the independent variable WP by presenting weather information using three different weather-information presentation symbologies. We refer to these as WP 1, WP 2, and WP 3. Beginning with the complete image for each variation, we created the set of 12 change trials and 8 catch trials—as described in Table 11 and Table 12—for each WP. WP was a between-subjects variable; each participant viewed trials from one of the three WPs.

3.1.5 Change-Detection Paradigm

To assess participants' ability to detect changes between two WP images (i.e., Image 1 and Image 2), we employed the *one-shot change-detection paradigm* described by Rensink (2002; see Figure 29). In a typical one-shot trial, Image 1 is displayed first for a period of several seconds. The display is then masked briefly by a blank screen, and then Image 2 is displayed. Image 2 remains on-screen until the participant presses one of two buttons to indicate that they detected a change (i.e., Image 2 was different than Image 1) or did not detect a change (i.e., Image 2 was the same as Image 1). A typical experiment using the one-shot paradigm includes both *change* and *no-change* (i.e., catch) trials.
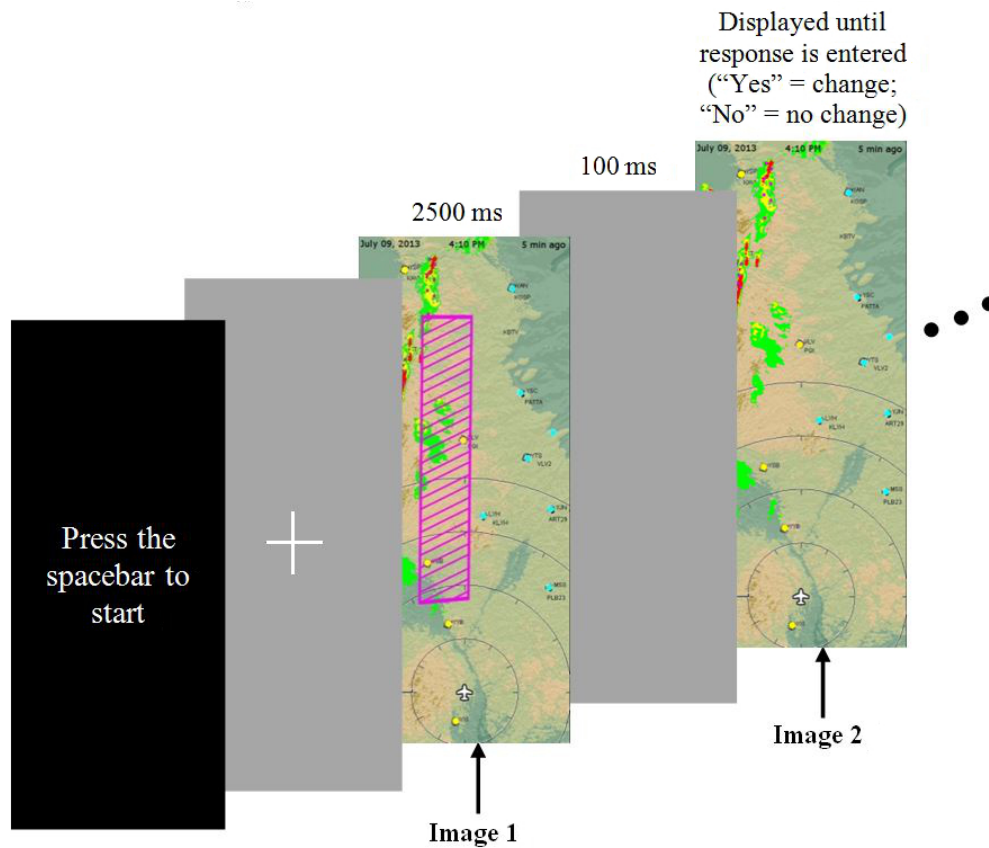

Figure 29. Illustration of the one-shot change-detection technique. Adapted from Rensink, 2002.

3.1.6 Stimulus Experiment System

The Stimulus Experiment System (SES), an in-house, custom-designed computer application, was used to present the change-detection trials and record participants' responses. Each change-detection trial comprised the following sequential displays (with display duration in parentheses):

- A white central fixation cross on a grey background (1,000 ms)

- Image 1 (2,500 ms)

- A blank, grey screen (i.e., interstimulus interval; 1,500 ms)

- Image 2 (remained on-screen until participants entered a response or for a maximum of 60,000 ms, whichever occurred first)

The display duration for Image 1 was determined after preliminary testing using durations of 2,500 ms and 5,000 ms. We observed a ceiling effect (i.e., near-perfect change-detection accuracy) using 5000 ms—but not 2,500 ms—and therefore we selected the shorter display duration. Similarly, the interstimulus intervals were determined after preliminary testing using durations of 100 ms, 1,000 ms, and 1,500 ms. We observed that increasing the interstimulus interval had the effect of reducing change-detection accuracy. For the experimental trials, we selected the longest (i.e., 1,500 ms) interstimulus interval because it is more representative of pilots' gaze behavior in the cockpit (i.e., pilots often look away from the weather display for more than 1,000 ms to fixate on other cockpit instruments and the OTW view, before refixating on the weather display), and because longer interstimulus intervals minimized the possibility of a ceiling effect.

3.2 Procedure

After participants completed an informed consent form and biographical questionnaire, the researcher used a randomized list to assign participants to one of the three weather-presentation variations, then started the SES and selected the appropriate variation, and then seated each participant at a computer. The researcher explained to the participant that the task instructions would be presented on-screen in a self-paced manner. The instructions emphasized that when responding during the change-detection trials, participants should prioritize accuracy over speed. After reading the on-screen instructions, participants first completed 14 practice trials (8 change trials, 6 no-change trials) followed by 60 experimental trials (12 change trials, each repeated three times; 8 no-change trials, each repeated three times). The experimental trials were presented in two blocks of 30 trials; trial order was randomized. Participants initiated each trial by pressing the spacebar on the keyboard; participants responded by pressing the key labeled *Yes* if they detected a change, or pressing the key labeled *No* if they did not detect a change. Participants could pause for as long as they desired after each trial and between the blocks (e.g., to drink, stretch, or use the restroom). Participants did not receive performance feedback (i.e., knowledge of results) during the practice or experimental trials.

3.3 Results and Conclusions

In this section, we report accuracy and response time results from the change-detection experiment. First, we report results from WP changes in location and color of METAR symbols. Subsequently, we report results from location changes to SIGMET areas, lightning symbols, precipitation symbols, and time-stamps. For all analyses, we only use the data for correct detection responses.

3.3.1 METAR Location Changes

The detection accuracy and the sensitivity to METAR location changes play a central role in the present study, and we based our power analysis on a research hypothesis for the detection accuracy of METAR location changes (see section Bayesian Power Analysis). In the real world, location changes to METAR symbols (i.e., the METAR is either present or absent as indicated by the filled or non-filled METAR symbol) could indicate that a ground reporting station is out of order or is not currently transmitting routine weather reports. Besides being absent or present, the METAR location changes also encompass a special case of METAR color changes. During these trials, the METAR symbol changes from no color (absent) to VFR color or from no color to IFR color. When the METAR symbols were present, there were 7 IFR symbols on the left side of the WP image and 7 VFR symbols on the right side of the WP image for a total of 14 METAR symbols. In the experiment, location changes were accomplished by using both onset (i.e., METAR symbol appearance) and offset (i.e., METAR symbol disappearance) trials for the METAR symbols as illustrated in Figure 30.
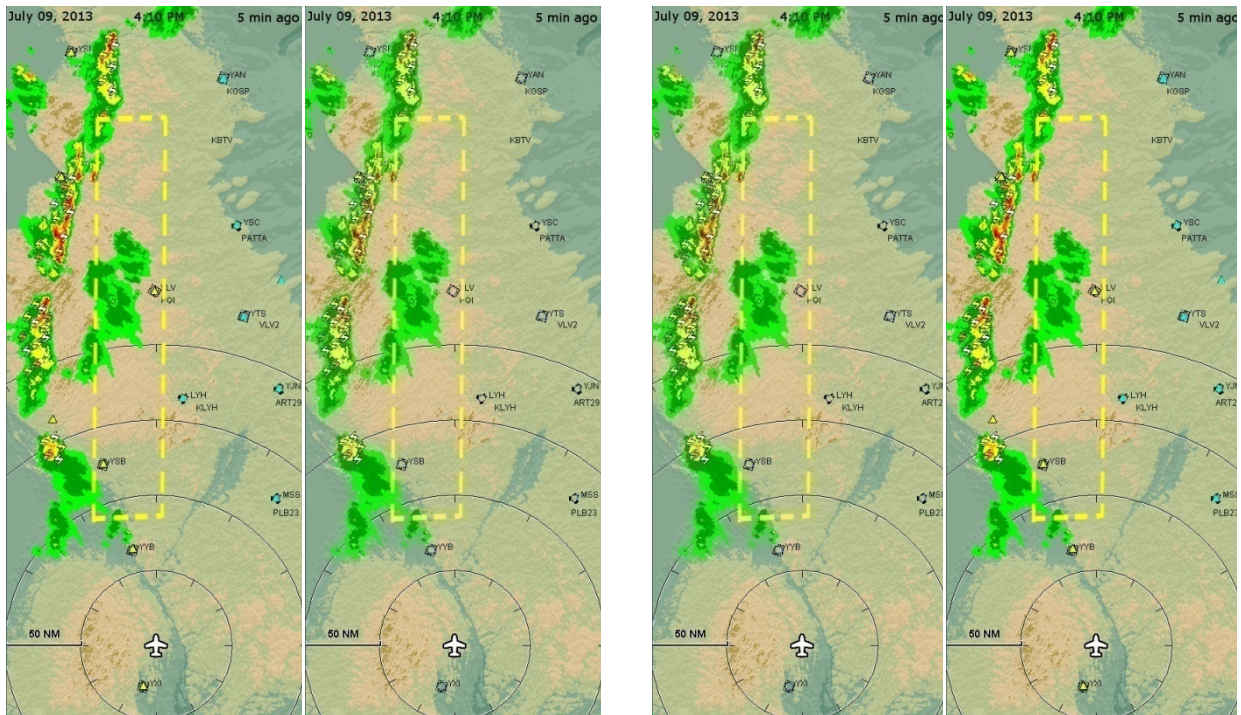

Figure 30. METAR offset (left) and onset (right) image pairs.

Figure 31 shows the METAR location change data (i.e., the individual pilot detection accuracy scores) for each of the three WPs (left). Figure 31 also shows the posterior distribution with μ_c (group means) and K_c (dispersion around μ_c) values for each WP. The posterior means for WPs 1-3 are .58, .79, and .84, respectively.
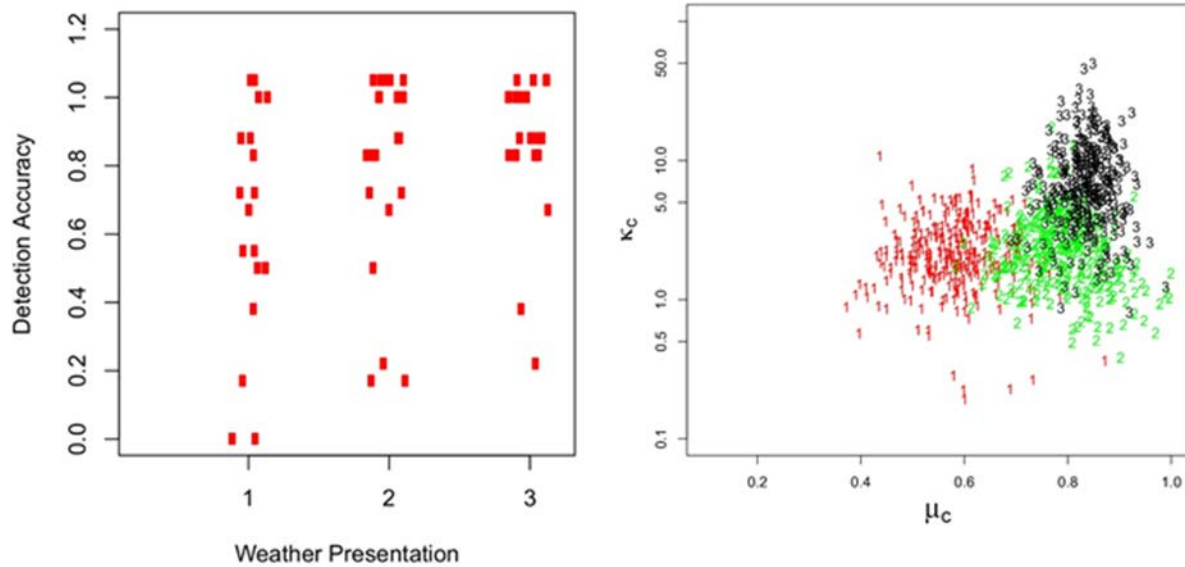

Figure 31. METAR detection data for the three WPs (left) and the posterior distribution (right). Note: We have perturbed each data score in the graph (left) to eliminate a complete overlap of data points. The detection accuracy score for each pilot is computed from the overall correct responses out of 6 trials. Therefore, each pilot can have an overall detection score of 0 (0 correct responses out of 6 trials), 0.16 (1 correct response out of 6 trials), 0.33 (2 out of 6), .5 (3 out of 6), .66 (4 out of 6), .83 (5 out of 6), or 1.0 (6 out of 6). The posterior distribution is presented as a scatter plot of μ_c (group mean) and K_c (dispersion of individual accuracy scores around the group mean) for each WP. During the analysis, we used 200,000 samples for the posterior. Only 300 of these samples are shown in the scatter plot to prevent clutter.

Figure 32 shows posterior contrasts for the comparison in detection accuracy between WPs 1-3. There is a credible difference between WP 1 and WP 2, with WP 2 having higher detection accuracy than WP 1. There is also a credible difference between WP 1 and WP 3, with WP 3 having higher detection accuracy than WP 1. However, there is no credible difference in detection accuracy between WP 2 and WP 3.
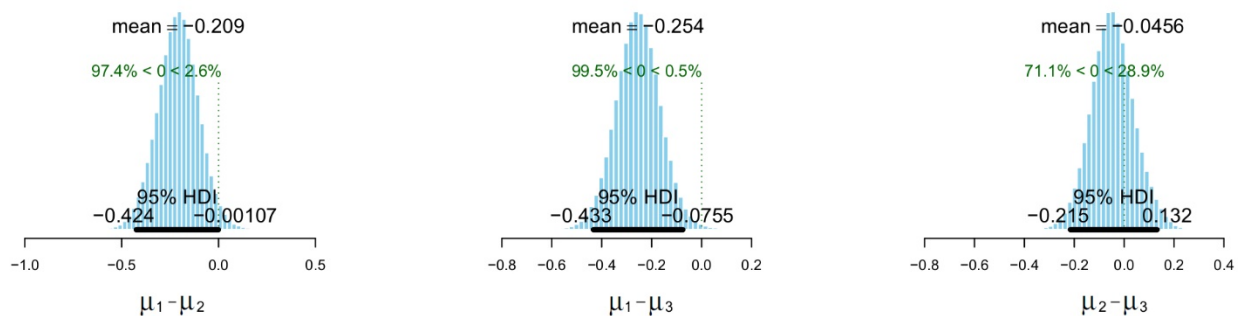

Figure 32. Posterior contrasts for the difference between WP 1 and 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

Because the three WP symbologies use different symbols to represent METARs we assessed whether triangles (WP 1) or circles (WP 2 and WP 3), on average, yielded the highest detection accuracy. We also assessed whether the METAR color combination blue/yellow (WP 1 and WP 2) or the color combination white/red (WP 3), on average, yielded the highest detection accuracy. Figure 33 shows the posterior contrasts for the triangles versus circles comparison (left) and the blue/yellow versus the white/red comparison. There is a credible difference in the detection accuracy between triangles and circles, with circles, on average, yielding higher detection performance than triangles. Also, there is a credible difference in detection accuracy between the two color versions with the white/red METAR symbol, on average, yielding higher detection accuracy than the blue/yellow METAR symbol.
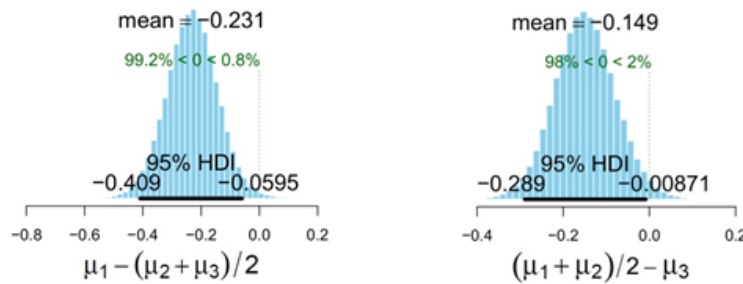
Figure 33. Posterior contrast for the difference in detection accuracy between METAR triangles and METAR circles (left), and the difference in detection between blue/yellow and white/red METAR symbols (right).

As part of our study goals, we stated three specific goals that relate to the detection accuracy of METAR location changes. We are mainly interested in comparing the mean detection accuracy for the three WPs, denoted by their group mean, μ. The first of our goals is that the mean of WP 2 exceeds the mean of WP 1, with the 95% HDI excluding the value 0 (i.e., $\mu1 - \mu2 > 0.0$). Our second goal is that the mean of WP 3 exceeds the mean of WP 1, with the 95% HDI excluding the value 0 (i.e., $\mu3 - \mu2 > 0.0$). Our third goals is that the mean of the two groups using circles (WP 2 and WP 3) exceeds the mean of the group using triangles (WP 1), with the 95% HDI excluding the value 0 (i.e., $\mu1 - (\mu2 + \mu3) / 2 > 0.0$). As we can see from the posterior contrasts in Figure 32 and Figure 33, we reached all three study goals.

We also analyzed the onset versus offset trials to assess if there are any differences in detection accuracy for trials where METAR symbols appeared versus disappeared in the WP image pairs. Figure 34 shows the posterior contrasts between the three WPS for the onset trials. Although there are no credible differences between WP 1 and WP 2, and WP 2 and WP 3, there is a credible difference between WP 1 and WP 3 with WP 3 having higher detection accuracy for onset trials than WP 1. There were no credible differences between the three WPs for offset trials; all 95% HDIs contained the value 0.
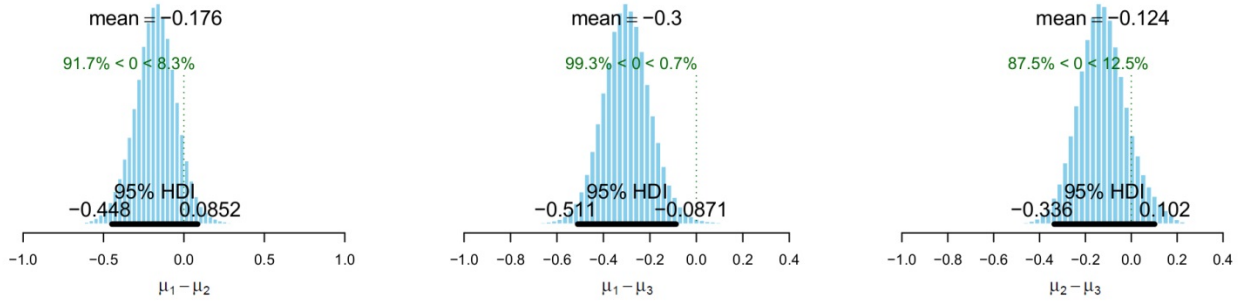
Figure 34. Posterior contrasts for the onset detection accuracy between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right) on onset trials.

In addition to the detection accuracy scores, we also recorded the response times for each trial during the experiment. For the response time analyses we used two response time values per pilot for each analysis; the average of the three onset trials and the average of the three offset trials. Figure 35 shows the response times for the METAR location changes.
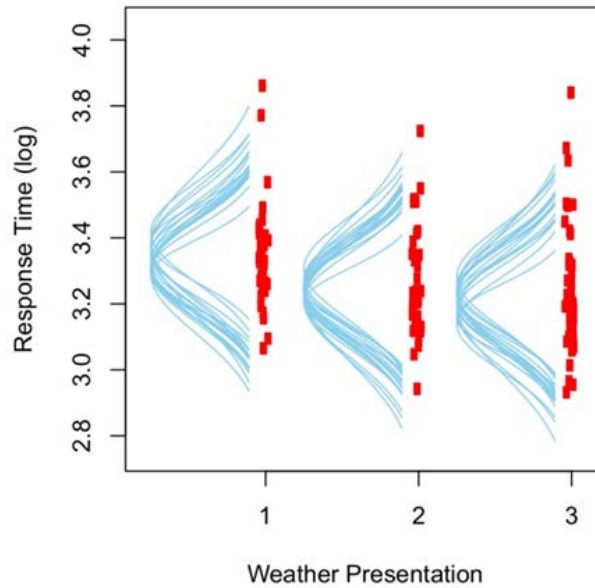
Figure 35. Response time data (log) for METAR location changes and posterior predictive check.

Figure 36 shows the posterior contrasts for the three WPs. There is a credible difference in the response times between WP 1 and WP 2, with longer response times for WP 1 than WP 2. Three is also a credible difference in response times between WP 1 and WP 3, with longer response times for WP 1 than WP 3. There is no credible difference between WP 2 and WP 3 because the value 0 is included in the 95% HDI.
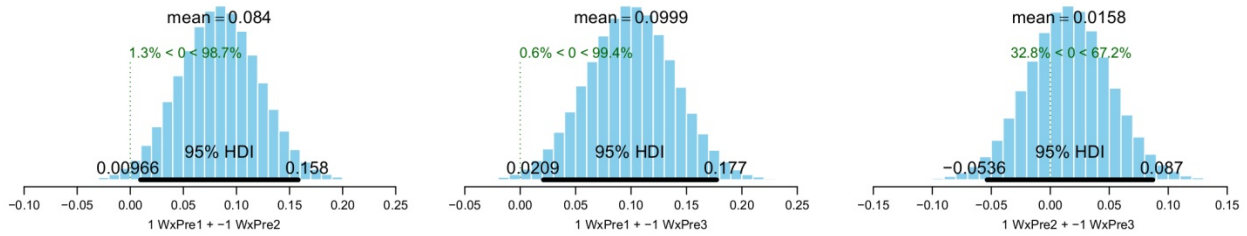

Figure 36. Posterior contrasts for METAR response times (log) between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

To sum up, the result shows credible differences between the three WPs in detection accuracy for METAR location changes. The detection accuracies for WP 2 (mean posterior detection accuracy, $\mu = .79$) and WP 3 (mean posterior detection accuracy, $\mu = .84$) are credibly higher than the detection accuracy for WP 1 (mean posterior detection accuracy, $\mu = .58$). Regarding METAR symbol shape and color, circles and white/red METAR colors yield higher detection performance, on average, than triangles and blue/yellow colors. Although the detection performance for onset versus offset trials is similar among the WPs, WP 3 yields credibly higher detection accuracy than WP 1 for onset trials.

3.3.2 METAR Color Changes

The color-coded METAR symbols indicate VFR or IFR flight conditions according to visibility and ceiling conditions at an airport. Of particular interest here is the detection of the change in METAR symbol colors from VFR to IFR, indicating a change from Visual Meteorological Conditions (VMC) to instrument meteorological conditions (IMC) at airports.

During the change-detection trials, the METAR symbols were all IFR color-coded in the first WP image and then changed to VFR color in the second WP image (IFR to VFR color change), or the METAR symbols were all indicating VFR in the first WP image and then appeared as IFR in the second image (VFR to IFR color change; see Figure 37).
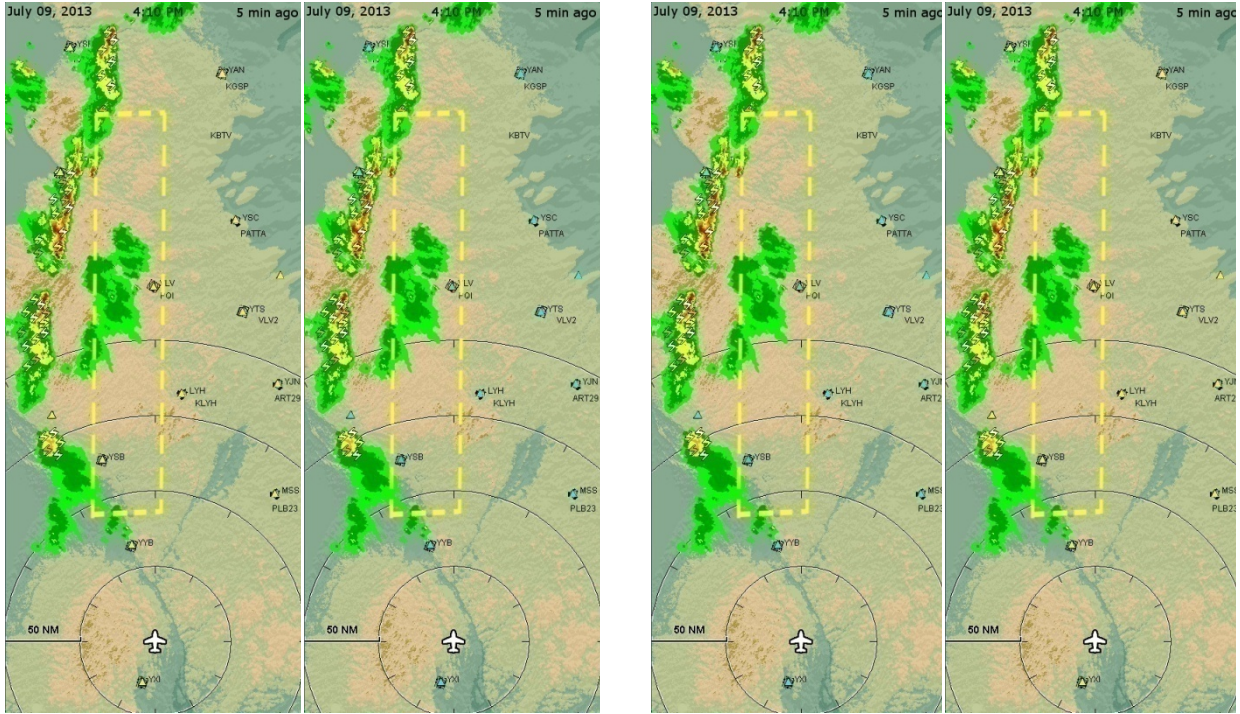
Figure 37. METAR color change images with IFR to VFR changes (left) and VFR to IFR changes (right).

Figure 38 shows the METAR location change data for each of the three WPs (left) and the posterior distribution with μ_c (group means) and K_c (dispersion around μ_c) values. The posterior means for WPs 1-3 are .60, .75, and .91, respectively.
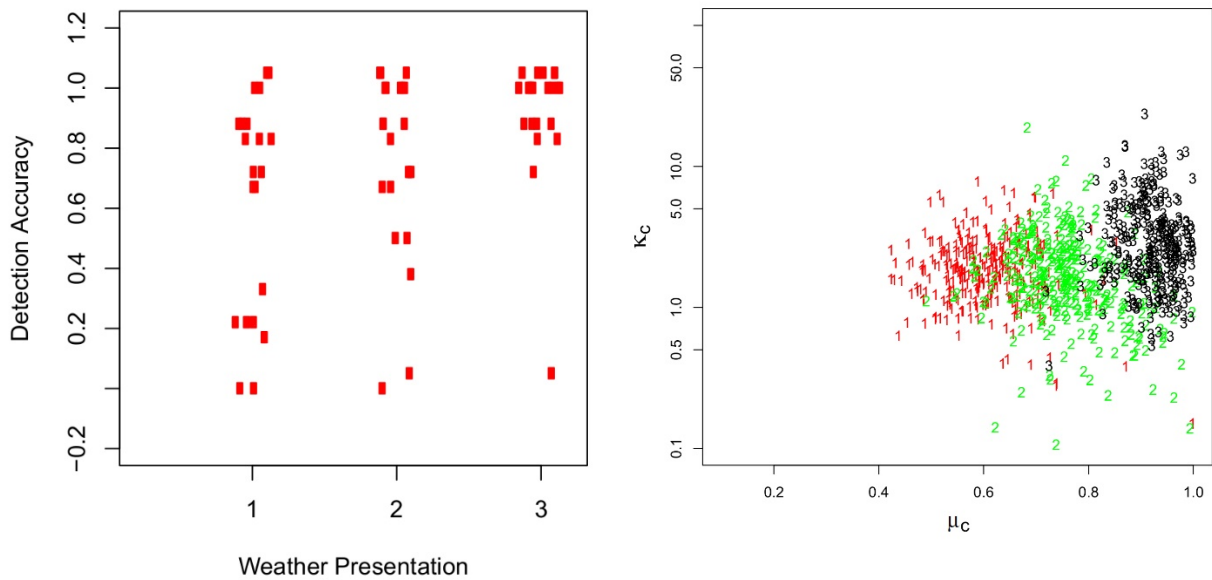

Figure 38. METAR color detection data (i.e., a color change from VFR to IFR, and from IFR to VFR) for the three WPs (left) and the posterior distribution (right).

Figure 39 shows posterior contrasts for the comparison in detection accuracy between WPs 1-3. Although there are no credible differences in detection accuracy between WP 1 and WP 2, and WP 2 and WP 3, there is a credible difference between WP 1 and WP 3 with WP 3 having higher detection accuracy than WP 1.
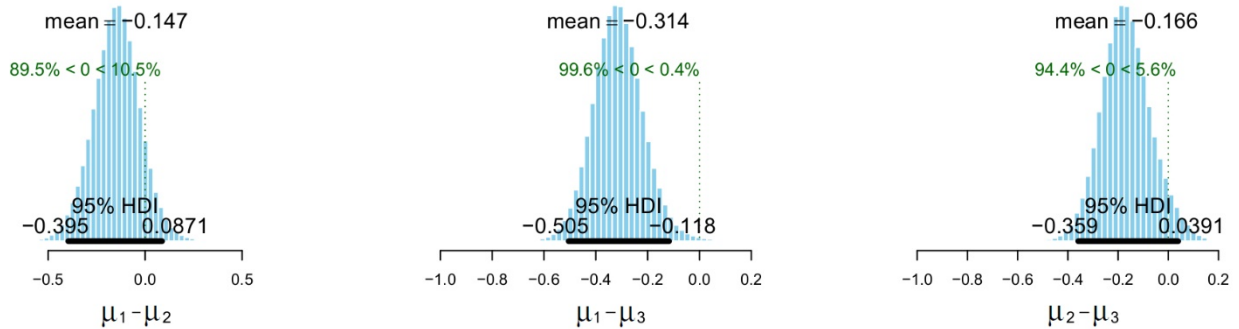
Figure 39. Posterior accuracy contrasts for the difference between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and 3 (right).

Figure 40 shows the posterior contrasts for the triangles versus circles comparison (left) and the blue/yellow versus the white/red comparison (right). There is a credible difference in the detection accuracy between triangles and circles, with circles, on average, yielding higher detection performance than triangles. There is also a credible difference in detection accuracy between the two color versions with the white/red METAR symbol, on average, yielding higher detection accuracy than the blue/yellow METAR symbol.
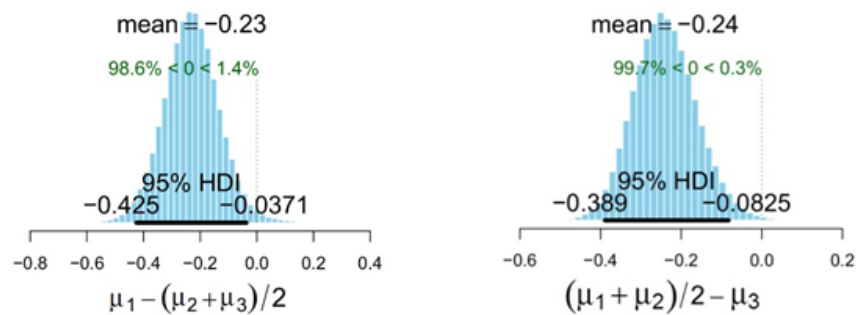

Figure 40. Posterior contrast for the difference in detection accuracy between METAR triangles and METAR circles (left), and the difference in detection between blue/yellow and white/red METAR symbols (right).

There were also differences between WPs with regards to detecting the METAR color change from VFR to IFR. Figure 41 shows that the detection performance for WP 3 is credibly higher than for WP 1. Figure 42 shows the response times for the detection of METAR color changes.
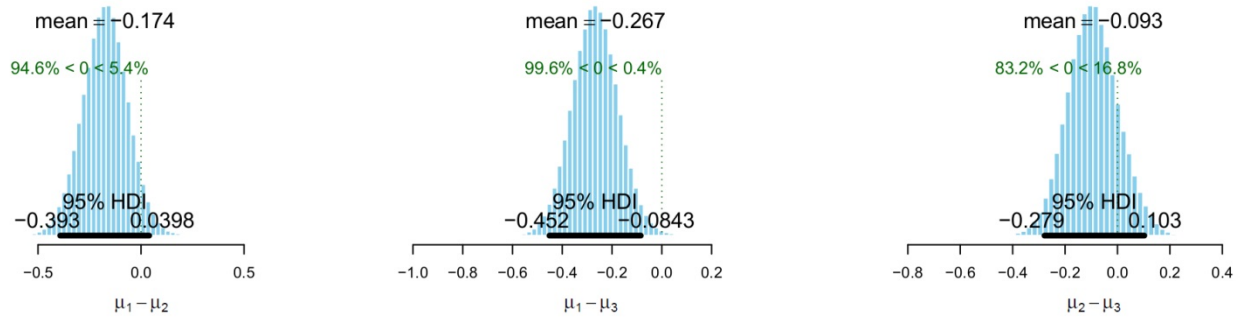

Figure 41. Posterior accuracy contrasts for VFR to IFR color changes between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).
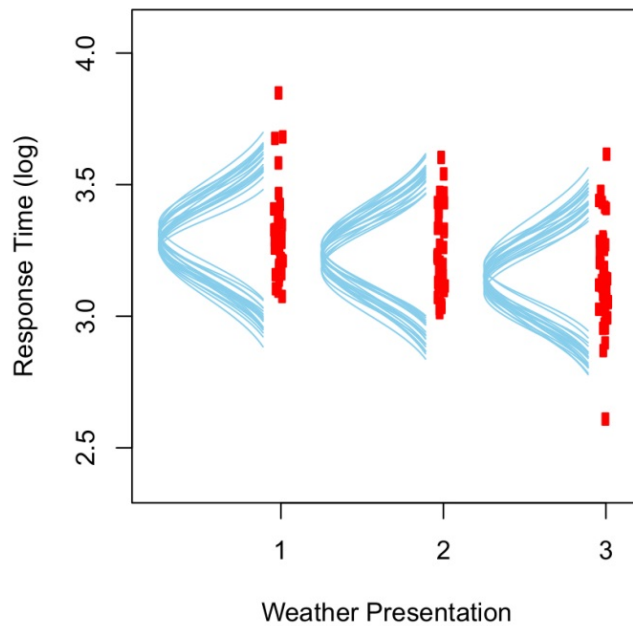

Figure 42. Response time data (log) for METAR color changes with posterior predictive check.

Figure 43 shows the posterior contrast for METAR color response times between WPs 1-3. Although there is no credible difference in response times between WP 1 and WP 2, there is a credible difference between WP 1 and WP 3 with WP 1 having longer response times than WP 3. There is also a credible difference between WP 2 and WP 3, with WP 2 having longer response times than WP 3.
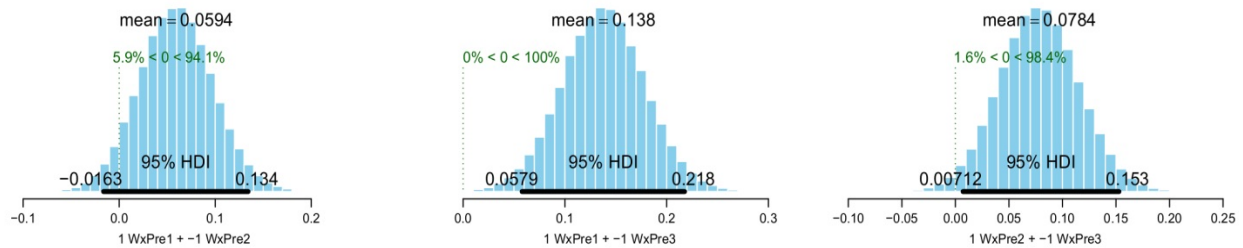

Figure 43. Posterior contrasts of METAR response times (log) between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

To sum up, there is a credible difference in detection accuracy of METAR color changes between WP 1 and WP 3, with WP 3 having higher detection accuracy than WP 1. On average, METAR circles yield higher detection performance than METAR triangles. The detection accuracy for white/red METAR symbols is, on average, higher than the accuracy for blue/yellow METAR symbols. There are also credible differences in response times with WP 1 having longer response times than WP 3, and WP 2 having longer response times than WP 3.

3.3.3 SIGMET Location Changes

During the change-detection trials, the SIGMET area was either present in the first WP image and then disappeared in the second WP image (offset trials), or the SIGMET area was absent in the first WP image and then appeared in the second image (onset trials) as illustrated in Figure 44.
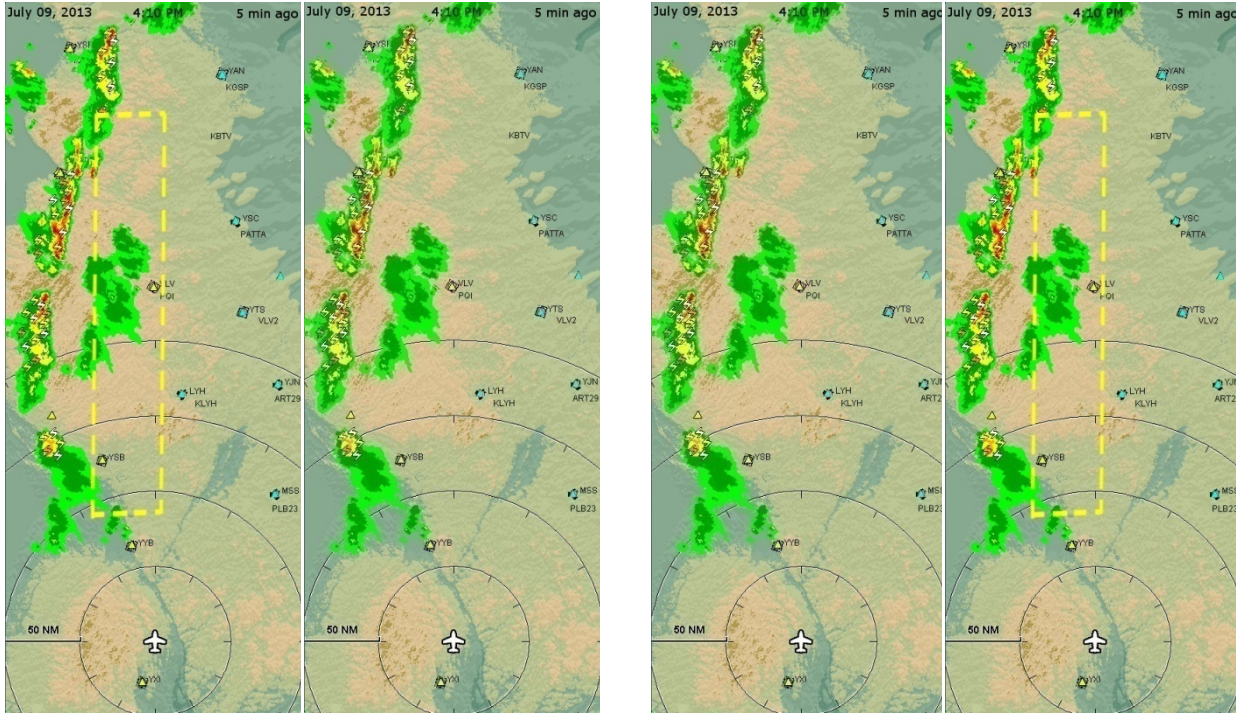
Figure 44. SIGMET offset (left) and onset (right) image pairs.

Figure 45 shows the SIGMET location change data for each of the three WPs (left) and the posterior distribution with μ_c (group means) and K_c (dispersion around μ_c) values. The posterior means for WPs 1-3 are .83, .93, and .86, respectively.
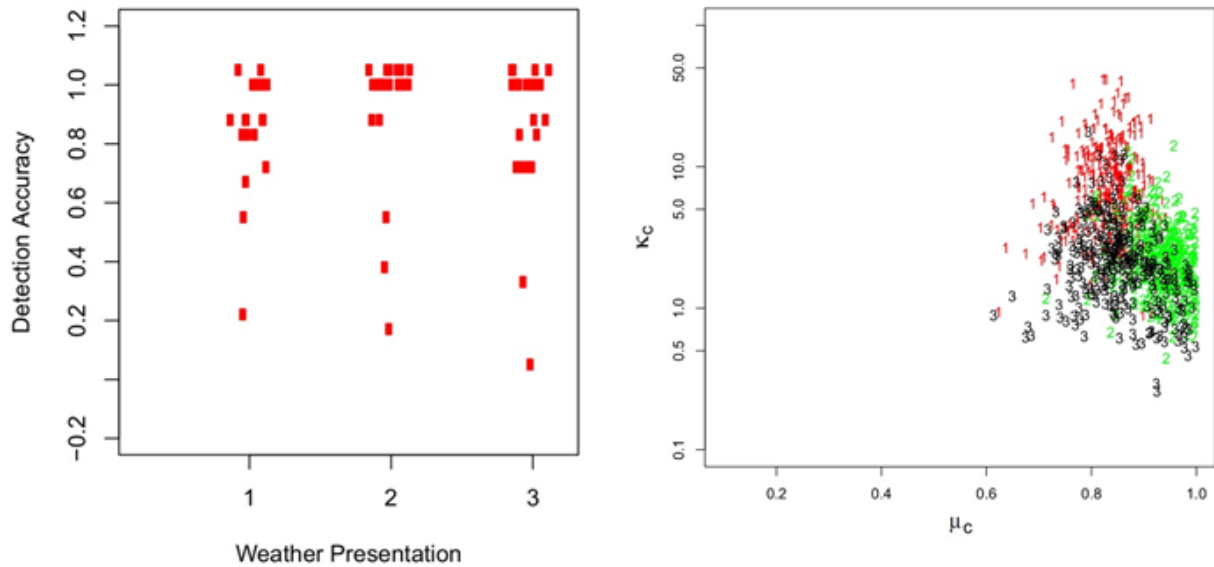

Figure 45. SIGMET detection data for the three WPs (left) and the posterior distribution (right).

56

With mean μ_c accuracies ranging from 83% to 93 %, detection performance was high on this task. As Figure 46 shows, there are no credible differences in detection accuracy between the three WPs. All posterior contrasts have the value 0 included in the 95% HDI.
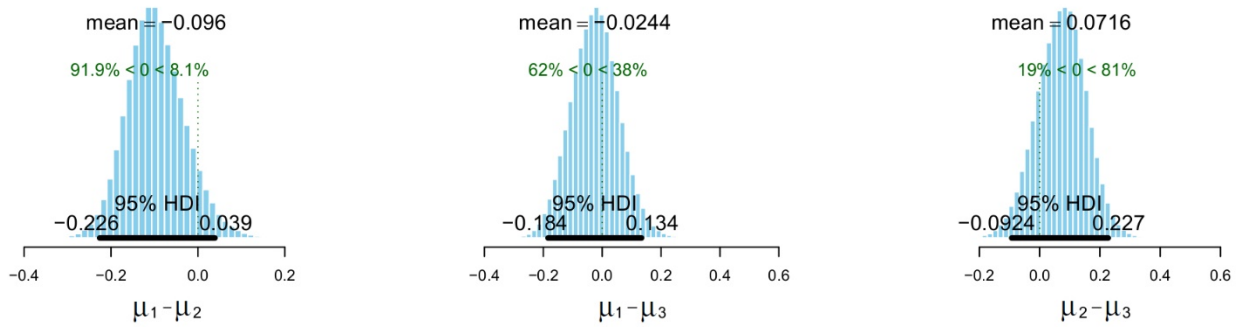

Figure 46. Posterior contrasts for the difference in SIGMET detection between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

Figure 47 shows the SIGMET response time data for WPs 1-3. Similar to the detection accuracy, there were no credible differences in response times among the three WPs. Figure 48 shows the response time contrasts for WPs 1-3, and all three HDIs include the value 0.
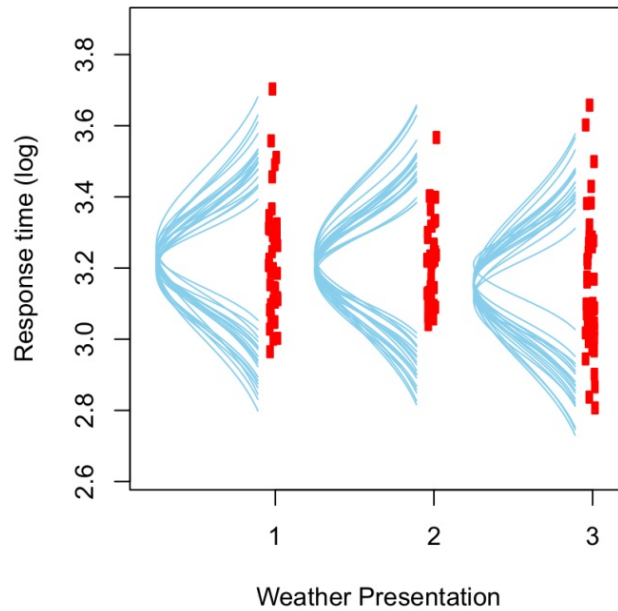

Figure 47. Response time data (log) for the detection of SIGMET location changes with posterior predictive check.
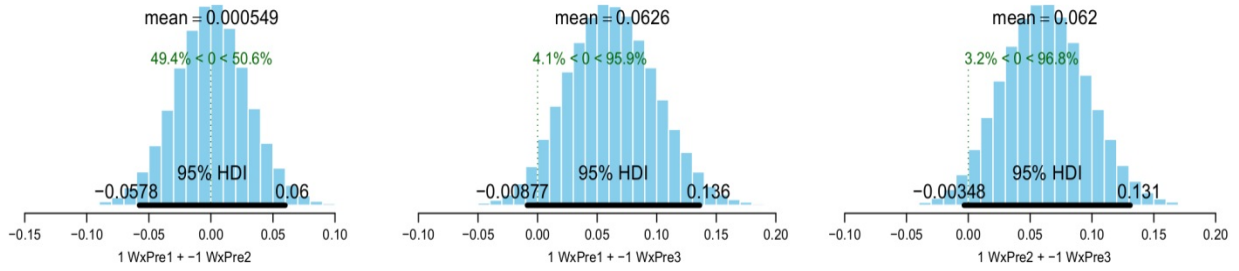
Figure 48. Posterior contrasts for SIGMET response times (log) between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right.

To sum up, there are no credible differences between WPs for the detection of SIGMET areas. Detection performance is high across all three WPs with predicted average correct detections ranging from 83% to 93%.

3.3.4 Lightning Location Changes

During the change-detection trials the lightning symbols were either present in the first WP image and then disappeared in the second WP image (offset trials), or they were absent in the first WP image and then appeared in the second image (onset trials). Figure 49 illustrates the lightning offset and onset image pairs.
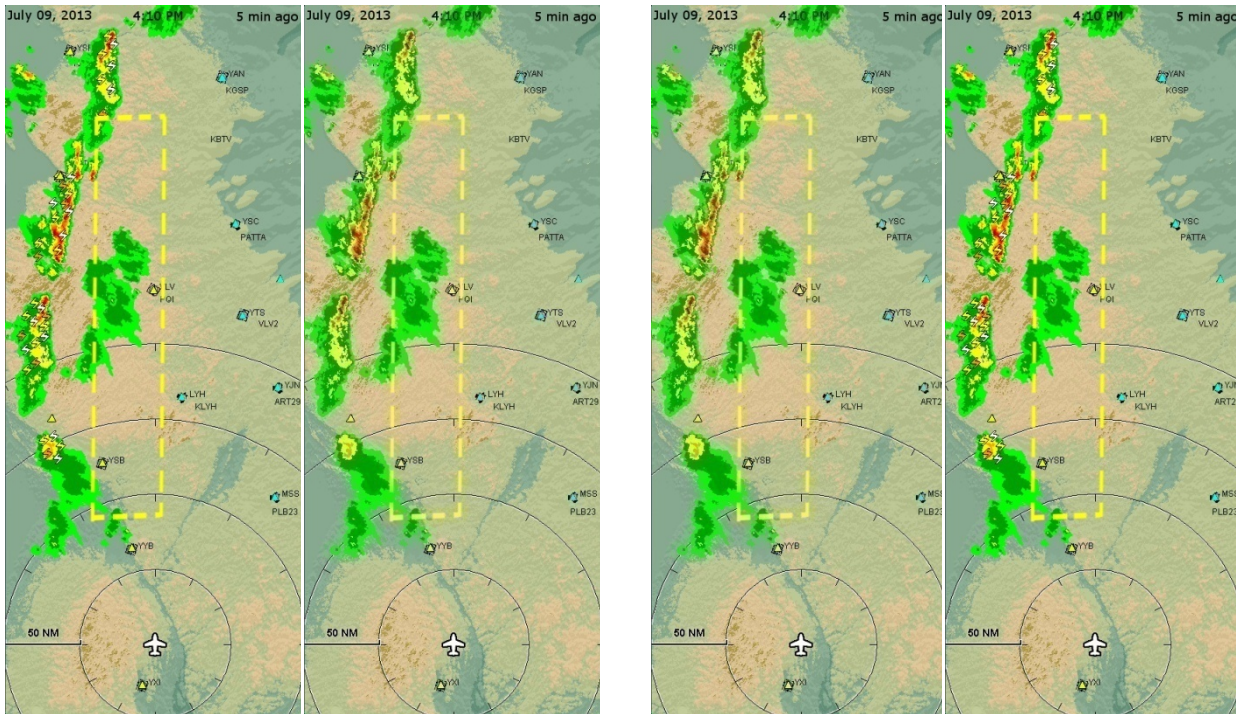

Figure 49. Lightning offset (left) and onset (right) image pairs.

Figure 50 shows the Lightning location change data for each of the three WPs (left) and the posterior distribution with μ_c (group means) and K_c (dispersion around μ_c) values. The posterior means for WPs 1-3 are .43, .20, and .18, respectively.
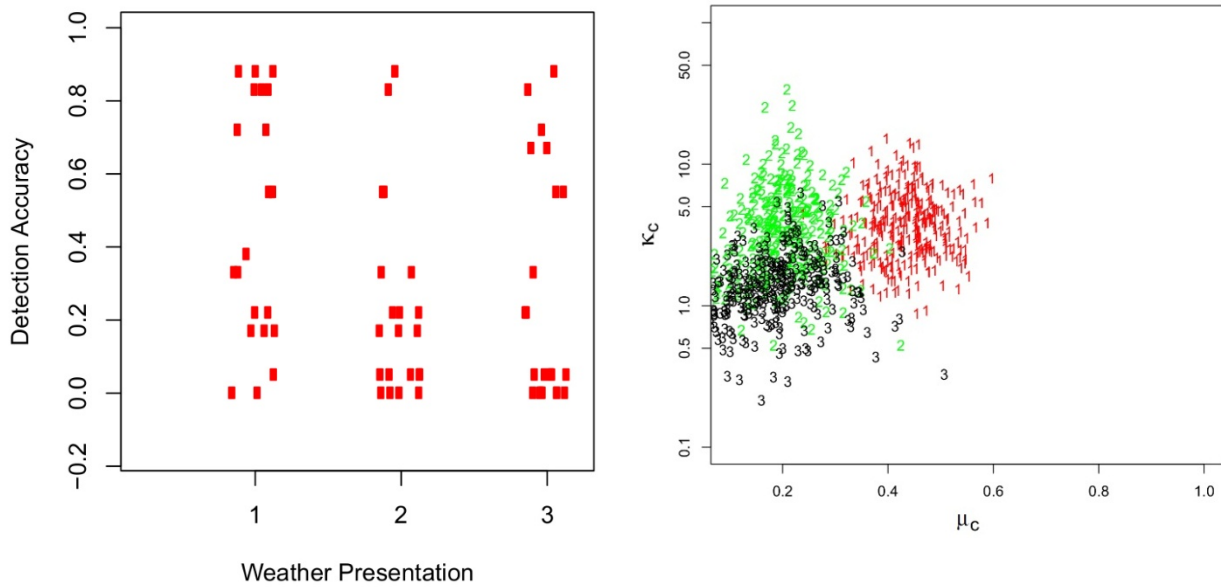

Figure 50. Lightning detection data for the three WPs (left) and the posterior distribution (right).

The accuracy for detecting a change in lightning positions was very low for all three WPs, with detection accuracies ranging from 18% to 43%. Nevertheless, there are differences in detection performance. Figure 51 shows the posterior contrast for WPs 1-3. There is a credible difference in detection performance between WP 1 and WP 2, with WP 1 having higher detection accuracy than WP 2. Also, there is a credible difference between WP 1 and WP 3, with WP 1 having higher detection accuracy than WP 3.
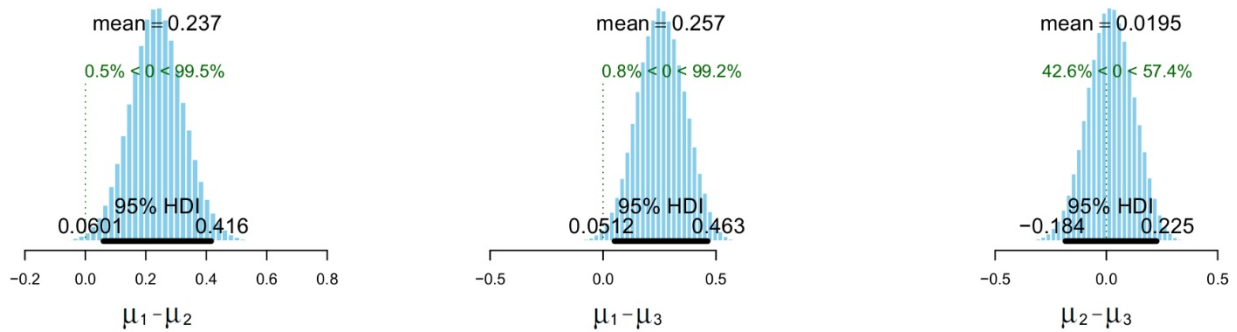

Figure 51. Posterior contrasts for the difference in lightning detection between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

For the detection of lightning location changes, there were also performance differences between WPs for the trials when lightning symbols appeared (onset trials) in one of the two images. Figure 52 shows the contrasts between WPs 1-3 detection accuracies for onset trials. There is a credible difference between WP 1 and WP 2, with higher detection accuracy for WP 1 compared to WP 2. There is also a credible difference in accuracy between WP 1 and WP 3, with WP 1 having higher accuracy than WP 3.
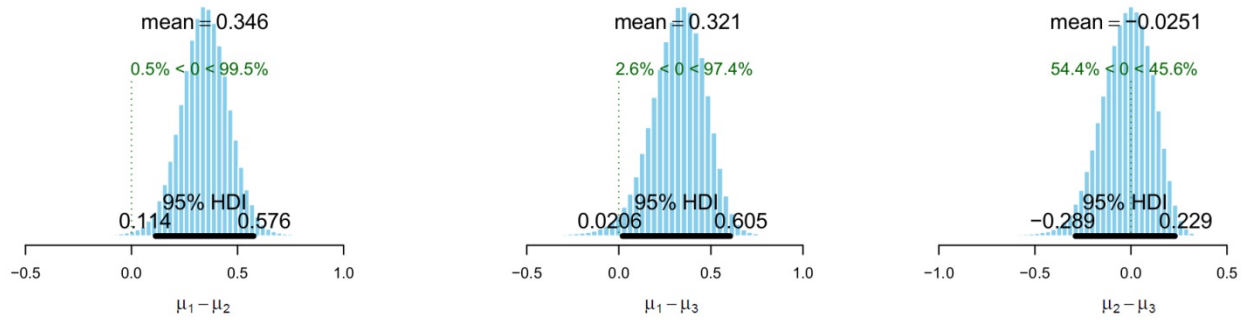

Figure 52. Posterior contrasts for the difference in onset detection accuracy between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

Figure 53 shows the response time data for WPs 1-3. There were no credible differences in response times. All posterior contrasts include the value 0 within the 95% HDI (see Figure 54).
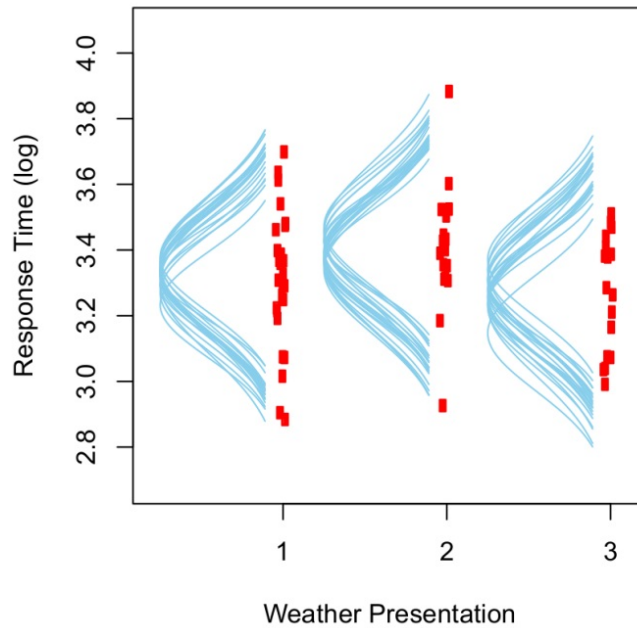

Figure 53. Response time data (log) for the detection of Lightning location changes with posterior predictive check.
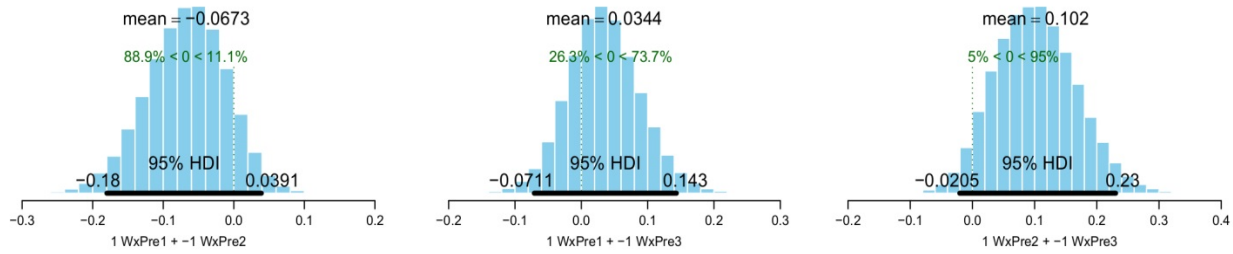
Figure 54. Posterior contrasts for Lightning response times (log) between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

To sum up, the performance for detecting changes in the location of lightning symbols is very low across the three WPs, with predicted average correct detections ranging from 18% to 43%. WP 1 yields higher detection accuracy than WP 2 and WP 3. Lightning symbols that portray a lightning bolt provide, on average, twice the predicted average detection accuracy (43%) compared to lightning symbols defined by a magenta *circle* (20%) or a yellow X (18%). Although the detection performance is similar across the three WPS for lightning symbols that *disappear* (offset) from the WP, there are credible differences between WPs in detection performance for lightning symbols that *appear* (onset) in a WP image. Although detection performance varies across the three WPs, there is no credible difference in response times.

3.3.5 Precipitation Location Changes

The precipitation location changes assessed pilot's sensitivity to the presence or absence of precipitation cells. During the change-detection trials, the precipitation cells were either present in the first WP image, and then disappeared in the second WP image (offset trials), or absent in the first WP image, and then appeared in the second image (onset trials). The precipitation onset and offset image pairs are illustrated in Figure 55.
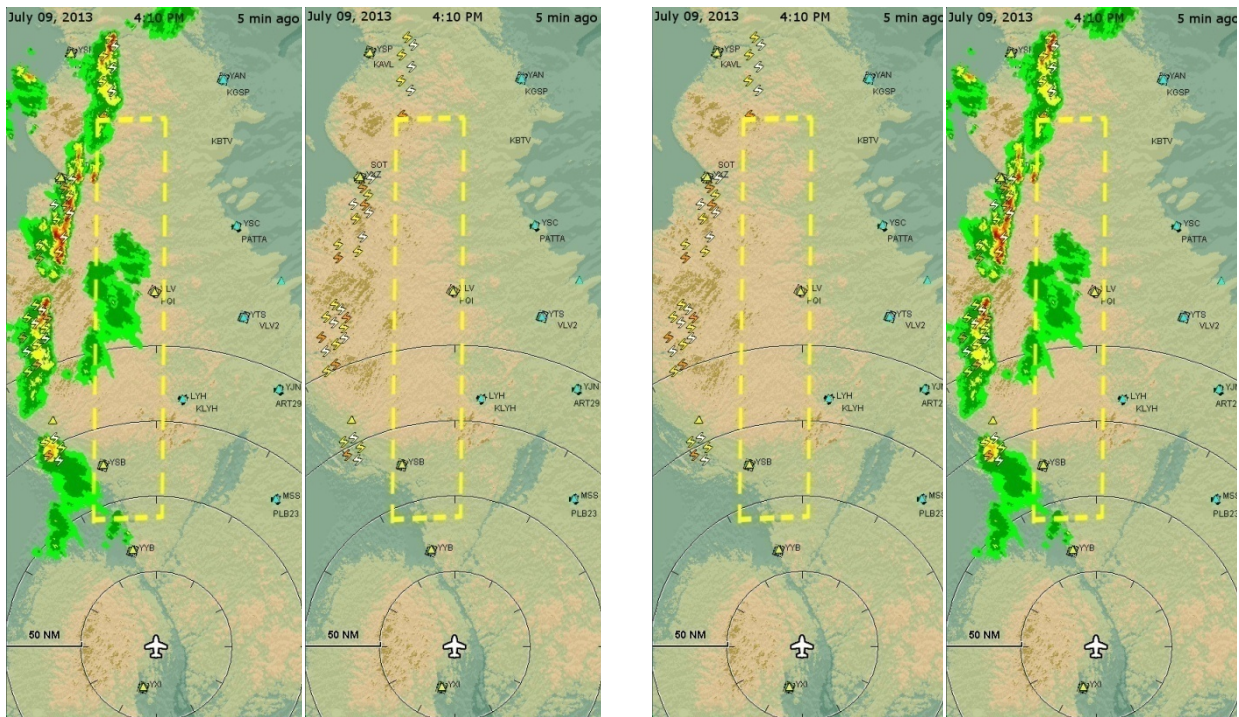
Figure 55. Precipitation offset (left) and onset (right) image pairs.

Figure 56 shows the Precipitation location change data for each of the three WPs (left) and the posterior distribution with μ_c (group means) and K_c (dispersion around μ_c) values. The posterior means for WPs 1-3 are .94, .89, and .91, respectively.
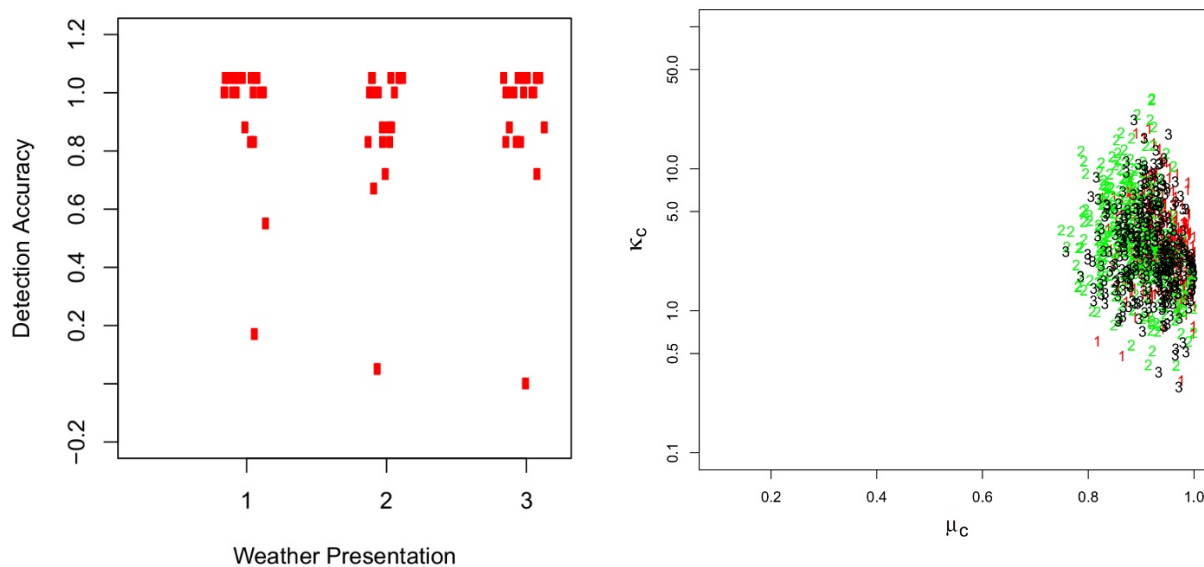

Figure 56. Precipitation detection accuracy data for the three WPs (left) and the posterior distribution (right).

The detection performance for precipitation location changes is high across all three WPs. There are no credible accuracy differences among the WPs. As Figure 57 shows, all posterior contrasts contain the value 0 with the 95% HDI.
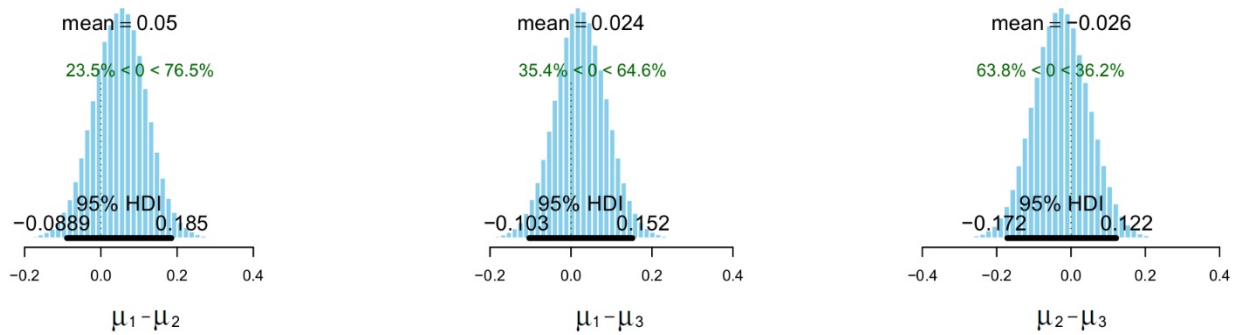

Figure 57. Posterior contrasts for the difference in precipitation detection between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

Figure 58 shows the response time data. There were no credible differences in response times; all posterior contrasts in Figure 59 contain the value 0 with the 95% HDI.
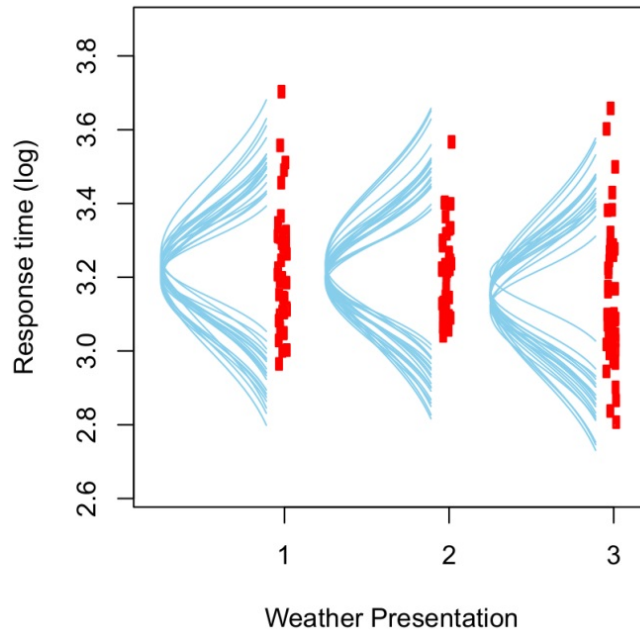

Figure 58. Response time data (log) for the detection of precipitation location changes with posterior predictive check.
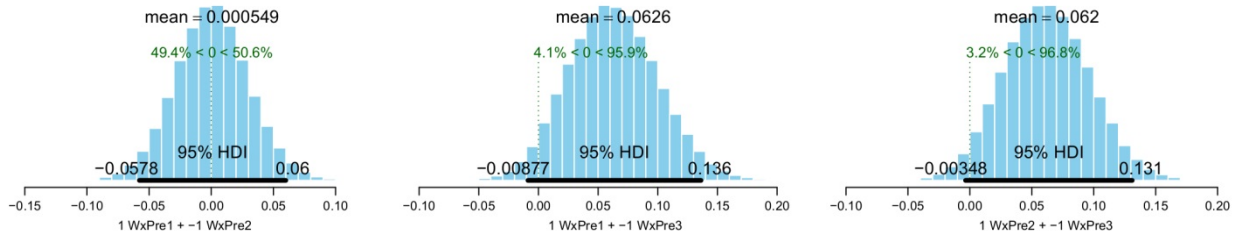
Figure 59. Posterior contrasts for precipitation response times (log) between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

To sum up, there are no credible differences between WPs for the detection of precipitation areas. Detection performance is high across all three WPs with predicted average correct detections ranging from 89% to 94%.

3.3.6 Time-stamp Location Changes

Commercially available WP products display time-stamp information in different ways and in some applications the user has to perform mental subtraction to derive the age of the weather data (FAA, 2010). In this study, we explore a simple time-stamp design that contains the date and time, and the duration (in minutes), since the weather display was last updated. The location of the time-stamp was always fixed at the top of the WP image, and the data within the time-stamp was always the same.

During the change-detection trials, the time-stamp either was present in the first WP image and then disappeared in the second WP image (offset trials) or was absent in the first WP image and then appeared in the second image (onset trials) as illustrated in Figure 60.
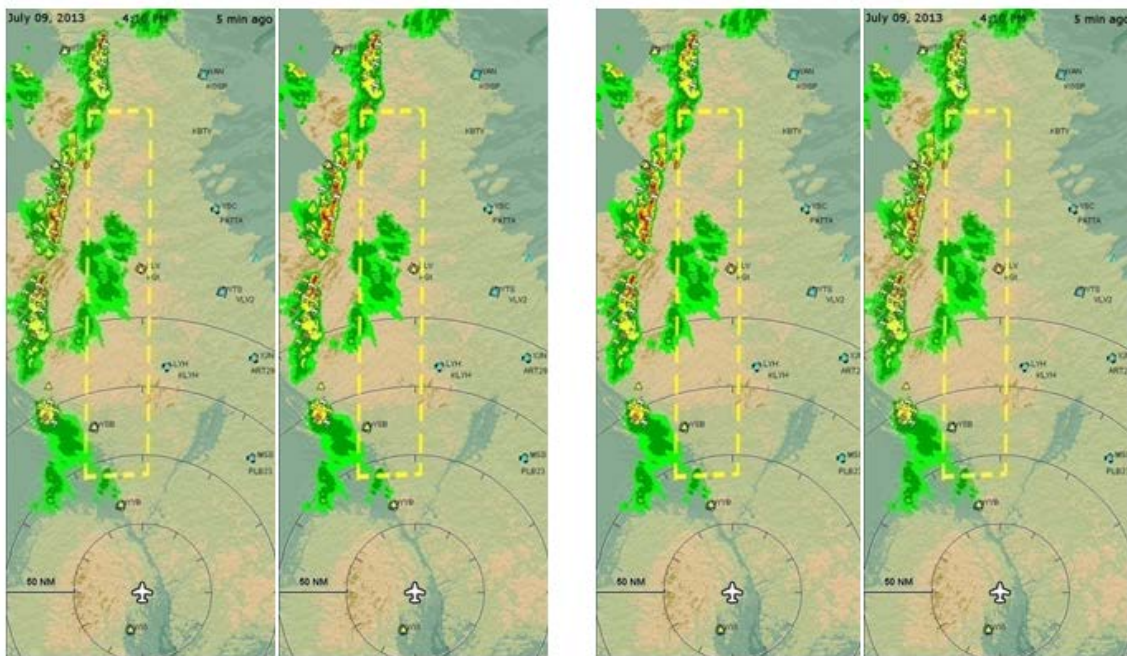

Figure 60. Time-stamp offset (left) and onset (right) image pairs.

Figure 61 shows the time-stamp location change data for each of the three WPs (left) and the posterior distribution with μ_c (group means) and K_c (dispersion around μ_c) values. The posterior means for WPs 1-3 are .20, .16, and .13, respectively.
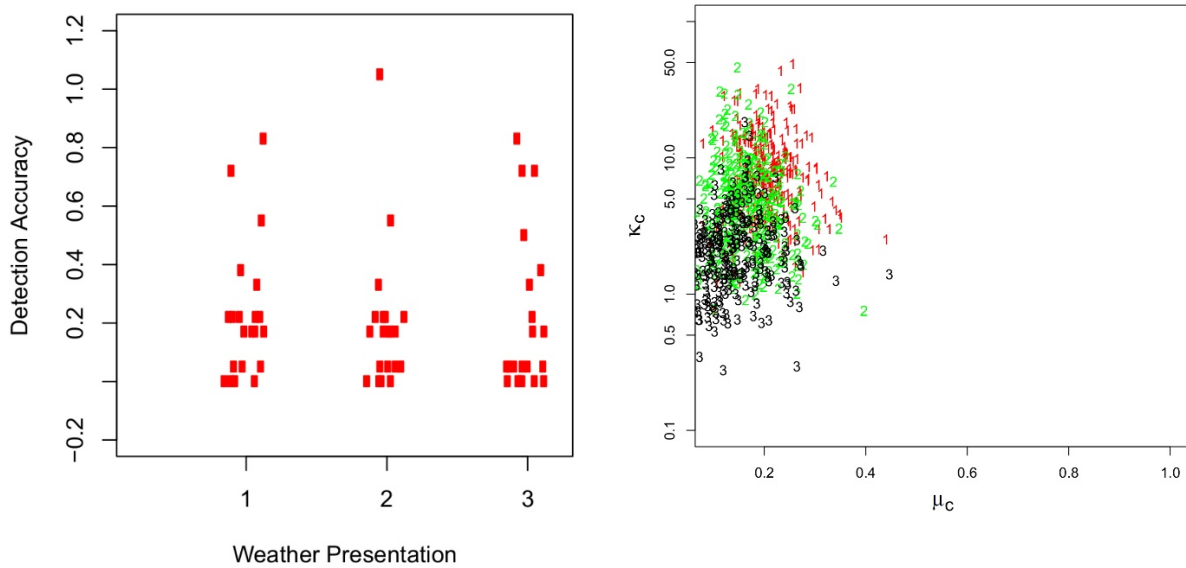

Figure 61. Time-stamp detection data for the three WPs (left) and the posterior distribution (right).

With mean predicted detection accuracies in the range of 13% to 20%, pilots were virtually blind to the time-stamp location changes. There are no credible differences in detection accuracy between the three WPs (see Figure 62).
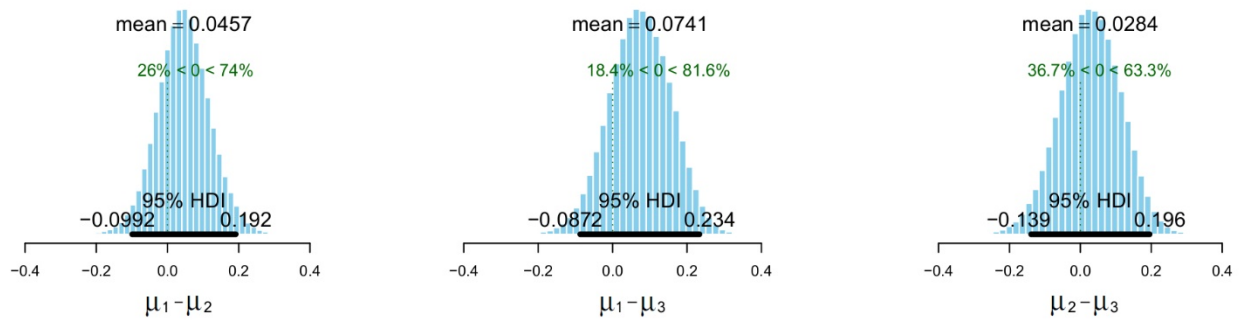

Figure 62. Posterior contrasts for the difference in time-stamp detection between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

Figure 63 shows the response time data for the time-stamp location changes. On average, the response times for the time-stamp images are longer than the response times for the remaining five weather elements used in the experiment.
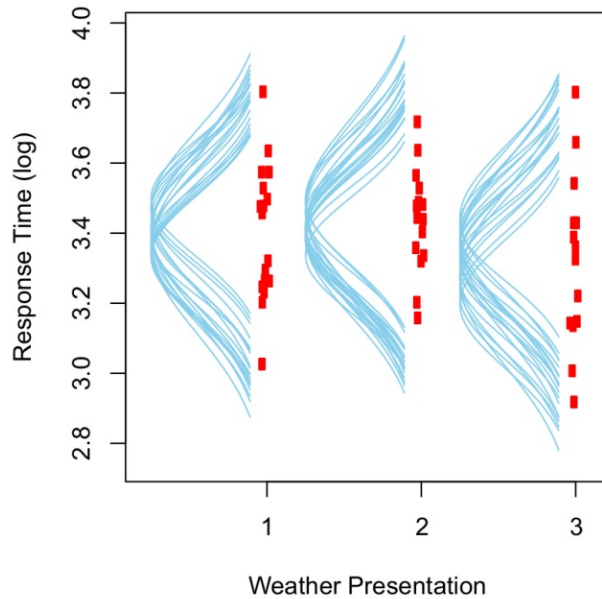

Figure 63. Response time data (log) for the detection of time-stamp location changes with posterior predictive check.

Figure 64shows the posterior contrasts for the time-stamp response times (log). There are no credible differences between WPs; all 95% HDIs include the value 0.
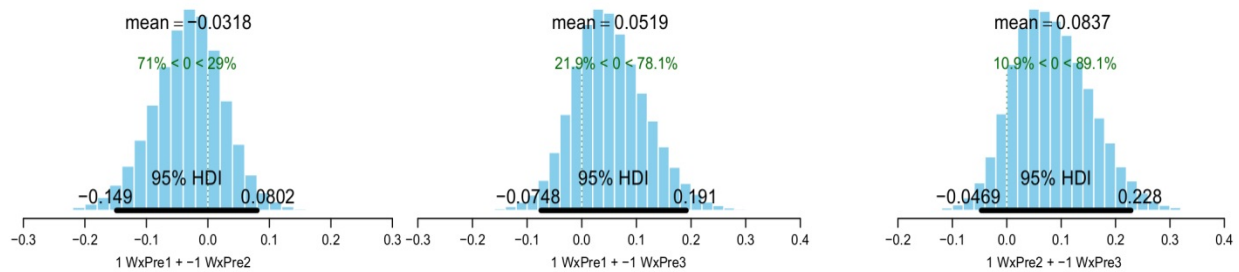

Figure 64. Posterior contrasts for time-stamp response times between WP 1 and WP 2 (left), WP 1 and WP 3 (middle), and WP 2 and WP 3 (right).

To sum up, there are no credible differences in detection accuracy between the three WPs for time-stamps. However, with mean predicted detection accuracies in the range of 13% to 20%, pilots are virtually blind to the presence or absence of time-stamps.

3.3.7 Retrospective Power Analysis

For our prospective power analysis we used data generated from a METAR research hypothesis. We ran repeated simulations and for each simulated experiment we checked the posterior distribution against our stated goals. The simulation outcome revealed 80% power in achieving our three goals using 20 participants per group.

Because we now have an actual posterior distribution from Experiment 2 we can repeat the power simulation process used for the prospective power analysis. But this time we are using our actual data and our actual posterior rather than an anticipated posterior derived from hypothesis-generated data. This retrospective power analysis adds nothing to our study inferences; we are simply curious and ask: How much power did we actually have?

The outcome from the retrospective power analysis revealed that we have 52% power in achieving Goal 1, $\mu1 - \mu2 > 0.0$; and 70% power in achieving Goal 2, $\mu3 - \mu2 > 0.0$ and Goal 3, $\mu1 - (\mu2+\mu3)/2 > 0.0$.

3.3.8 Replication Probability

Another analysis of interest concerns replication probability. We would like to know our probability of exactly replicating the current outcome, if we were to collect data from a new sample of pilots running the exact same experiment. For this power simulation we take our current data into account, effectively using our actual posterior from our actual data as the prior for the new simulated data sets. The outcome of the replication probability simulation shows that we have a 78% replication probability in achieving Goal 1 (i.e., mean detection accuracy of WP 2 exceeds the mean detection accuracy of WP 1) and a 95% replication probability in achieving Goal 2 (i.e., mean detection accuracy of WP 3 exceeds the mean detection accuracy of WP 1), and a 94% replication probability in achieving Goal 3 (i.e., mean detection accuracy from METAR circle symbols exceed the mean detection accuracy from METAR triangle symbols).

3.4 Discussion

In this study, we assessed pilots' perception of weather symbology changes. During a simulated flight, pilots navigated a pre-planned route from VOR to VOR while performing common pilot tasks—such as *see and avoid* (during VFR), reading charts, operating radio and navigational frequencies, listening to radio communications, viewing approach plates, and observing the cockpit instruments and the weather presentation (WP). While performing these tasks, pilots typically allocate their focus of attention to distinct cockpit areas corresponding to the OTW view, the glass instrument display, the WP, the console, and the sectional map (Ahlstrom & Dworsky, 2012). In the course of pilots' multitasking, we introduced METAR-symbol changes that signaled reduced ceiling and visibility conditions (i.e., VFR to IFR) at selected airports. Our main interest was to assess pilot perception of symbol change and to assess whether the perception of change was the same for pilots using different WPs.

The result shows that pilots (using different WPs) vary considerably in their overall perception of METAR symbol change during flight. The overall group detection performance ranges from a virtual blindness (25% detections) to a modest detection performance (62% detections). However, because these results are from a simulated flight where pilots are multitasking while piloting, there are many uncontrolled variables that might affect the perception of change. Therefore, we needed to isolate the detection task in a change-detection experiment (Experiment 2) to see how pilots perform when they focus their visual attention solely on detecting symbol changes.

Furthermore, while the simulated flight (Experiment 1) focused on METAR changes only, in Experiment 2, we wanted to include additional symbols to assess the detectability for each symbol included in the three WPs. The result from the change-detection experiment shows that the detection accuracy varies greatly between different weather symbols and between different WPs. Although the average change-detection performance is high across all WPs for precipitation areas

(on average, 89% to 94% correct detections), SIGMET areas (83% to 93%), and METAR symbols (83% to 91%), pilots are virtually blind to changes for lightning symbols (17% to 43%) and time-stamp information (13% to 20%).

This outcome clearly shows that WP symbology affects pilots' perception of symbol change and cognitive engagement. Pilot performance varies credibly between different symbology renderings of the same weather data. Although this is a negative outcome considering the vast number of available WP symbologies, it is important empirical information that can help us develop more optimal presentations (e.g., symbol shapes and chromaticity: Ahlstrom & Arend, 2005; Arend, 2003). Preferably, WPs should display symbols that allow rapid encoding and detection. This is especially important considering the large number of different weather elements that can be overlaid on modern multifunctional displays using different backgrounds (FAA, 2010). As more symbols and background areas are color-coded, the possible combinations of foreground and background colors rapidly increase. This can lead to salience problems where more important information (e.g., METAR-symbol color change) fails to visually segregate from less critical background information. We need presentation symbologies that achieve good margins of legibility and detectability for all combinations of symbols and background colors.

Although it is central to have optimal weather symbology for all aviation users, it is especially important for single-pilot flights where the purpose of the WP is to allow the pilot to continuously update his or her weather situation awareness. Piloting requires multitasking, and multitasking requires divided visual attention. When used for pre-flight planning, we have a different situation because none of the time constraints and the divided attention associated with piloting are present. During flight, the main concern is to make sure that pilots perceive, and are aware of, any changes to weather symbols. In this situation, the increased number of weather symbols and the complexity of visual layers will likely work against a pilot. For exploratory use of weather information during pre-flight planning, pilots are likely to benefit from an increased number of weather elements and visual overlays as it allows the pilot to explore different "what-if" scenarios in areas relevant to the intended route of flight.

It is clear from this study that pilots' perception of symbol changes while in flight is frail, leaving many changes undetected. This change blindness is a well-known phenomenon (Rensink, 2000, 2002) that is particularly strong during multitasking situations (Varakin, Levin, & Fidler, 2004). If WP symbols are conveying essential and flight-relevant information, and it is important that pilots perceive changes to this information, then there needs to be a presentation method that provides a connection between the presentation and the pilot. This could be accomplished in various ways, for example, through alerts or through algorithms that keep track of the weather information and notifies the user (Ahlstrom & Jaggard, 2010). Army researchers using the Force XXI Battle Command, Brigade, and Below Display (Durlach, 2004) have also found evidence for operator change blindness. The Army researchers found, among other things, considerable change blindness for color changes where participants detected only 63.9% of all the display changes. This result is similar to the best overall METAR detection accuracy from the simulation flights in the present study. Because of the resilience of the change blindness phenomena, the Army researchers expressed concerns that improvements in display symbology might not be sufficient to remedy this problem. Instead, they proposed that other aids—such as intelligent alerts and event logs (or change database)—might be required to make sure users perceive new display changes and that they are aware of previous changes (Durlach & Meliza, 2004). Although a change database or event log would not be suitable for single-pilot flights, it could be of use as an option to display historical weather information during pre-flight planning.

The change blindness phenomenon works against a pilot in many different ways. In the present study, failure to notice symbol changes led to some undesirable consequences. For example, pilots who failed to detect the initial METAR change were more likely to continue their VFR flight towards the pre-planned destination without good weather situation awareness. Had these pilots been aware of the initial METAR change at their destination airport (signaling reduced ceiling and visibility) they would likely have considered requesting weather updates from ATC, considered an alternate destination airport, or requested an IFR flight plan. A failure to detect the METAR changes leads to time and space compression. Pilots end up with a reduced time span for decision-making as they get closer to the intended destination without good weather situation awareness, with fewer alternate destinations prepared, and sometimes without the possibility to land. Being aware of weather changes as early as possible is advantageous because it allows the pilot more time to evaluate information and to make guided decisions. This is especially important for VFR flights when the pilot cannot fly in IMC. Granted, all pilots in the present simulation were IFR rated and equipped to fly in IMC. Therefore, pilots might not have perceived that they were in need of additional weather information or saw a need to request an IFR flight plan. However, a failure to detect the METAR changes and not requesting weather updates from ATC can be unfavorable even during IFR. For example, while the ceiling and visibility conditions at a destination airport legally allow a pilot to land, the situation might nonetheless be below a pilot's personal minima and therefore prevent the pilot from landing.

To sum up, weather information updates while in flight could potentially assist VFR pilots in avoiding IMC. Modern electronic cockpit displays and hand-held devices use graphical symbols to represent weather information elements. Pilots need to monitor these weather presentations and be tuned to symbol changes to maintain their weather situation awareness. In a simple world, *what* weather information is presented to pilots would matter only, not *how* it is presented. But as the present study shows, it is how it is presented to pilots that matters. Not every symbol is a good one and not every combination of symbols and colors produce ideal or even equivalent presentations. In a multitasking cockpit environment, these effects work against pilots to maintain their weather situation awareness. Symbols update their location and change colors, but pilots often cannot detect the changes. Therefore, the time has come to direct efforts for the development of weather presentations that not only present weather information but also ascertain that pilots see it and are aware of the updated information.

References

Ahlstrom, U., & Arend, L. (2005). Color usability on air traffic control displays. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 93-97). Santa Monica, CA: Human Factors and Ergonomics Society.

Ahlstrom, U., & Dworsky, M. (2012). *Effects of weather presentation symbology on general aviation pilot behavior, workload, and visual scanning* (DOT/FAA/TC-12/55). Atlantic City International Airport, NJ: FAA William Hughes Technical Center.

Ahlstrom, U., & Jaggard, E. (2010). Automatic Identification of Risky Weather Objects in Line of Flight (AIRWOLF). *Transportation Research Part C: Emerging Technologies, 18,* 187–192.

Arend, L. (2003). Graphics issues of an aviation integrated hazard displays. *Proceedings of the International Symposium on Aviation Psychology, 12,* 60–64.

Ball, F., Elzemann, A., & Busch, N. A. (2013). The scene and the unseen: Manipulating photographs for experiments on change blindness and scene memory. *Behavior Research Methods.* Advance online publication.

Beringer, D., & Ball, J. (2004). *The effects of NEXRAD graphical data resolution and direct weather viewing on pilots' judgments of weather severity and their willingness to continue flight* (DOT/FAA/AM-04/5). Oklahoma City, OK: Civil Aerospace Medical Institute Federal Aviation Administration.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378–399.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research, 3*(1), 58–67.

Coyne, J. T., Baldwin, C. L., & Latorella, K. A. (2005). Influence of graphical METARs on pilot's weather judgment. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 131–135). Santa Monica, CA: Human Factors and Ergonomics Society.

Durlach, P. J. (2004). Army digital systems and vulnerability to change blindness. In H. Kwon, N. M. Nasrabadi, W. Lee, P. D. Gader, & J. N. Wilson (Eds.), *Proceedings of the Twenty-Fourth Army Science Conference* (Accession No. ADM001736). Redstone Arsenal, AL: Army Missile Research, Development and Engineering Lab.

Durlach, P. J., & Meliza, L. L. (2004). The need for intelligent change alerting in complex monitoring and control systems.In J. Gunderson & C. Martin (Eds.), *Interaction between humans and autonomous systems over extended operation*: *papers from the aaai spring symposium* (Technical Report No. SS-04-03, pp. 93–97). Menlo Park, CA: AAAI Press.

Elgin, P. D., & Thomas, R. P. (2004). *An integrated decision-making model for categorizing weather products and decision aids.* Hampton, VA: NASA.

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

Federal Aviation Administration. (2010). *Weather technology in the cockpit program capabilities report* (DTFAWA-09-C-00088). Norman, OK: Atmospheric Technology Services Company, LLC.

Federal Aviation Administration. (2014). *FAR/AIM* [Federal aviation regulations/Aeronautical information manual]. New York, NY: FAA.

Federal Aviation Administration, & National Oceanic and Atmospheric Administration. (2010). *Aviation weather services* (AC 00-45). Oklahoma City, OK: FAA and NOAA.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587–606.

Goodman, S. (2008). A dirty dozen: Twelve *p*-value misconceptions. *Seminars in Hematology, 45*, 135–140.

Goodman, S. N. (1992). A comment on replication, *p*-values and evidence. *Statistics in Medicine, 11*, 875–879.

Grasse, T., Schilke, C., & Schiefele, J. (2008). Symbology evaluation for strategic weather information on the flight deck. *Proceedings of the IEEE/AIAA Digital Avionics Systems Conference, 27*, 4.B.1-1–4.B.1-12.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician, 57*, 171–182.

Ison, D. (2014, January/February). Understanding VFR into IMC accidents. *Plane & Pilot*, 1–3.

Izzetoglu, K., Bunce, S. C., Shewokis, P. A., & Ayaz, H. (2010). *Conformance monitoring and controller workload part task* (DTFA01-00-C-00068). Philadelphia , PA: Drexel University Press.

Izzetoglu, M., Bunce, S. C., Izzetoglu, K., Onaral, B., & Pourrezaei, K. (2007). Functional brain imaging using near-infrared technology: Assessing cognitive activity in real-life situations. *IEEE Engineering in Medicine and Biology Magazine*, *26*(4), 38–46.

Johnson, N., Wiegmann, D., & Wickens, C. (2006). Effects of advanced cockpit displays on general aviation pilots' decision to continue visual flight rules flight into instrument meteorological conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50,* 30–34.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 658–676.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier. doi:101016/jtics201005001

Latorella, K. A., & Chamberlain, J. P. (2002a). *Graphical weather information system evaluation: Usability, perceived utility, and preferences from General Aviation pilots* (NASA-2002-01-1521). Hampton, VA: NASA.

Latorella, K. A., & Chamberlain, J. P. (2002b). Tactical vs. strategic behavior: General aviation piloting in convective weather scenarios. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 46*, 101–105.

McAdaragh, R. M. (2002). *Toward a concept of operations for aviation weather information implementation in the evolving national airspace system* (NASA/TM-2002-212141). Hampton, VA: NASA.

McDougall, S. J. P., de Bruijn, O., & Curry, M. B. (2000). Exploring the effects of icon characteristics on user performance: The role of icon concreteness, complexity, and distinctiveness. *Journal of Experimental Psychology: Applied, 6*, 291–306.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Retrieved from http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf

Plummer, M. (2011). *RJAGS: Bayesian graphical models using MCMC. R package version 3-5* [Computer software]. Retrieved from http://CRAN.R-project.org/package=rjags

R Development Core Team. (2011). *R: A language and environment for statistical computing* [Computer software manual]. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rensink, R. A. (2000). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition, 7*, 345–376.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53*, 245–277.

RTCA. (2004). *Minimum aviation system performance standards (MASPS) for flight information service-broadcast (FIS-B) data link* (DO-267A). Washington, DC: RTCA.

The Stan Development Team. (2013). *Stan: A C++ library for probability and sampling* [Version 2.2.0]. http://mc-stan.org/

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Varakin, D. A., Levin, D. T., & Fidler, R. (2004). Unseen and unaware: Implications of recent research on failures of visual awareness for human–computer interface design. *Human-Computer Interaction, 19*, 389–422.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779–804.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*, 426–432.

Yuchnovicz, D. E., Novacek, P. F., Burgess, M. A., Heck, M. L., & Stokes, A. F. (2001). *Use of data-linked weather information display and effects on pilot navigation decision making in a piloted simulation study* (NASA/CR-2001-211047). Hampton, VA: NASA.

Acronyms

ATC	Air Traffic Control
BANOVA	Bayesian Analysis of Variance
DUATS	Direct User Access Terminal System
ETX	East Texas VOR
FAA	Federal Aviation Administration
fNIR	Functional Near Infrared Spectroscopy
GA	General Aviation
GIMP	GNU Image Manipulation Program
GWIS	Graphical Weather Information System
HDI	High Density Interval
HSI	Horizontal Situation Indicator
IFR	Instrument Flight Rules
IMC	Instrument Meteorological Conditions
KABE	Allentown, Pennsylvania Airport
KMRB	Martinsburg, West Virginia Airport
METAR	Meteorological Report
NAS	National Airspace System
NEXRAD	Next Generation Radar
NextGen	Next Generation Air Transportation System
NHST	Null Hypothesis Significance Testing
NOAA	National Oceanic and Atmospheric Administration
OTW	Out-The-Window
PTT	Push-To-Talk
SAGAT	Situation Awareness Global Assessment Technique
SIGMET	Significant Meteorological Advisory
SME	Subject Matter Expert
VFR	Visual Flight Rules
VMC	Visual Meteorological Conditions
VOR	Omnidirectional Radio Range
WJHTC	William J. Hughes Technical Center
WP	Weather Presentation

Appendix A: Biographical Questionnaire

Biographical Questionnaire

Demographic Information and Experience

1. What pilot certificate and ratings do you hold? (circle as many as apply)	Private Commercial ATP Glider SEL SEA MEL Airship Instrument CFI CFII MEI Helicopter A&P IA

2. What is your age?	_____ **Years**

3. Approximately, what is your total time?	_____ **Hours**

4. Approximately how many actual instrument hours do you have?	_____ **Hours**

5. Approximately how many instrument hours have you logged in the last 6 months (simulated and actual)?	_____ **Hours**

6. List all (if any) in-flight weather presentation systems you have used during a flight to make actual weather judgments (not including onboard radar or Stormscope).

_____

_____

_____

7. Have you had any training in weather interpretation other than basic pilot training (for example, courses in meteorology)? If so, to what extent?

_____

_____

_____

Thank you very much for participating in our study, we appreciate your help.

Appendix B: Weather Briefing

Weather Briefing

Condensed FAA Direct Access User Terminal System (DUATS) Weather Briefing Information
Date: June 9, 2011

SYNOPSIS AND VFR CLOUDS/WEATHER FORECASTS
BOSC FA 120946 CORRECTION
SYNOPSIS AND VFR CLOUDS/WEATHER
SYNOPSIS VALID UNTIL 130400
CLOUDS/WEATHER VALID UNTIL xx2200...OUTLOOK VALID xx2200-130400
NJ PA WV MD DC.

SEE AIRMET SIERRA FOR IFR CONDITIONS AND MTN OBSCURATION.
THUNDERSTORM IMPLY SEVERE OR GTR TURBULENCE SEVERE ICE LOW LEVEL
WIND SHEAR AND IFR CONDITIONS.
NON MSL HEIGHTS DENOTED BY ABOVE GROUND LEVEL OR CEILING.

PA NJ
NORTHWESTERN PA..BROKEN 5/8-7/8 COVERAGE AT 4000 FT TOP 070. 12Z
SCATTERED 3/8-4/8 COVERAGE AT 5000 FT. OUTLOOK..VFR.
SOUTHWESTERN PA-N CENTRAL PA..BROKEN 5/8-7/8 COVERAGE AT 2000 FT
LAYERED FL200.
VIS 3SM SCATTERED HEAVY THUNDERSTORM(S). 19ZBROKEN 5/8-7/8 COVERAGE
AT 3500 FT TOP 070. SCATTERED LIGHT SHOWER(S) OF RAIN 18Z BROKEN 5/8-7/8
COVERAGE AT 5000 FT. OUTLOOK..VFR.
S CENTRAL PA-NERN PA..BROKEN 5/8-7/8 COVERAGE AT 2500 FT LAYERED FL220.
VIS 10SM SCATTERED 1215 SCATTERED 3/8-4/8 COVERAGE AT 5000 FT.
OUTLOOK..VFR.
SOUTHEASTERN PA-NRN NJ..BROKEN 5/8-7/8 COVERAGE AT 1500 FT LAYERED
FL220. VIS 3SM LIGHT SNOW MIST. 15Z BROKEN 5/8-7/8 COVERAGE AT 8000 FT TOP
140. OUTLOOK..VFR.
SOUTHERN NJ..BROKEN 5/8-7/8 COVERAGE AT 10000 FT TOP FL200. SCATTERED
LIGHT SHOWER(S) OF RAIN. BECOMING 1214 BROKEN 5/8-7/8 COVERAGE AT 1000
FT LAYERED FL220. VIS 3SM LIGHT RAIN.18Z BROKEN 5/8-7/8 COVERAGE AT 7000
FT TOP 150. OUTLOOK..VFR

WV MD DC DE VA
WESTERN WV..BROKEN 5/8-7/8 COVERAGE AT 2000 FT LAYERED FL200. VIS 3SM
SCATTERED
BECOMING 1517 BROKEN 5/8-7/8 COVERAGE AT 2500 FT TOP 050.
OUTLOOK..MARGINAL VFR CEILING..01Z VFR.
N MOUNTAINS WV-E WV PNHDL-MD PANHANDLE..OVERCAST AT 3500 FT TOP 120.
VIS 3SM SCATTERED LIGHT SHOWER(S).
OUTLOOK..MARGINAL VFR CEILING.
SOUTHEASTERN WV-SWRN VA..BROKEN 5/8-7/8 COVERAGE AT 6500 FT TOP 100.
VIS 5SM SCATTERED LIGHT SHOWER. 14Z OVERCAST AT 4000 FT.
OUTLOOK..VFR.......20Z MARGINAL VFR FOG MIST

SURFACE WEATHER OBSERVATIONS
KABE (ALLENTOWN, PA) SCHEDULED OBSERVATION 1734UTC,
   WIND FROM 330 DEGREES AT 06 KTS, GUSTING TO 11 KTS,
   VISIBILITY 10.00 MILES,
   SKY BROKEN 5/8-7/8 COVERAGE AT 6000 FT, OVERCAST AT 7000 FT,
   TEMPERATURE 1C (33 DEG F), DEW POINT -7C (20 DEG F),
ALTIMETER SETTING 30.16 INCHES.
   REMARKS: AO2 SNE02B37E45 SLP216 P0000 T00061067
KABE (ALLENTOWN, PA) SCHEDULED OBSERVATION 12/2100 UTC,
   WIND FROM 330 DEGREES AT 11 KTS, GUSTING TO 15 KTS,
   VISIBILITY 10.00 MILES,
   SKY BROKEN 5/8-7/8 COVERAGE AT 6000 FT, OVERCAST AT 7000 FT,
   TEMPERATURE 1C (33 DEG F), DEW POINT -7C (20 DEG F),
ALTIMETER SETTING 30.16 INCHES.
KMRB (MARTINSBURG, WV) SCHEDULED OBSERVATION 12/1900 UTC,
   WIND FROM 330 DEGREES AT 06 KTS, GUSTING TO 11 KTS,
   VISIBILITY 10.00 MILES,
   SKY SCATTERED 3/8-4/8 COVERAGE AT 8500FT,
   TEMPERATURE 2C (35 DEG F), DEW POINT -3C (27 DEG F),
ALTIMETER SETTING 30.22 INCHES.
   REMARKS: AO2   RAB00E21UPB11E20 SLP237 P0000 T00221033

===>FORECAST CONDITIONS<===

   TERMINAL FORECASTS
KABE (ALLENTOWN, PA) AERODROME FORECAST AMENDED 12/1325 UTC,
   FOR USE FROM xx1300Z TO 131200Z,
   AT 121300Z, WIND FROM 330 DEGREES AT 12 KTS, GUSTING TO 20 KTS,
   VISIBILITY OVER 6.00 MILES, SKY SCATTERED 3/8-4/8 COVERAGE AT 3000 FT,
   OVERCAST AT 5000 FT,
   TEMPORARY CHANGES BETWEEN 121300Z AND 121500Z, VISIBILITY 5.00 MILES,
   WEATHER HEAVY THUNDERSTORMS, MIST, SKY OVERCAST AT 2500 FT,
   FROM 121600Z, WIND FROM 330 DEGREES AT 14 KTS, GUSTING TO 22 KTS,
   VISIBILITY OVER 6.00 MILES, SKY OVERCAST AT 5000 FT,
   FROM 122300Z, WIND FROM 310 DEGREES AT 10 KTS, VISIBILITY OVER 6.00
   MILES,
   SKY BROKEN 5/8-7/8 COVERAGE AT 25000 FT,

KMRB (MARTINSBURG, WV) AERODROME FORECAST 12/1800 UTC,
   FOR USE FROM xx1200Z TO xx1200Z,
   AT xx1500Z, WIND FROM 330 DEGREES AT 8 KTS, VISIBILITY
   OVER 6 MILES, SKY OVERCAST AT 5000 FT,
   FROM xx2000Z, WIND CALM, VISIBILITY 1 MILES FOG MIST,
                        SKY SCATTERED 3/8-4/8 COVERAGE AT 5000 FT,

Appendix C: Probe Questions

Probe Questions

Probe Questions Administered at the 11, 20, and 35-Minute Marks of the Simulated Flight

Note: The questions designed to assess whether participants detected each METAR change are bolded.

$t = 11$ min (At $t = 10$ min the METAR at KMRB changed from VFR to IFR)

1. Have you checked in with ATC? If so, which ATC control facility did you first check in with?
2. At what altitude did you check in with ATC?
3. What ATC control facility were you handed off to?
4. After the East Texas VOR, what was your next navigational facility and what course did you set?
5. **Were there any thunderstorms or other weather-related changes in the areas of Dulles and Martinsburg?**

$t = 20$ min (At $t = 19$ min the METARs at KBWI, KDCA, KESN, KIAD, and KJST changed from VFR to IFR)

1. What ATC control facility are you communicating with?
2. What navigational facility are you using?
3. What is your current heading?
4. **Did you notice any changes in the on-screen weather presentation since the last time the simulation was paused?**

$t = 35$ min (At $t = 30$ min the METAR at KHGS changed from VFR to IFR)

1. What ATC control facility are you communicating with?
2. Did you notice any changes in the on-screen weather presentation since the last time the simulation was paused?
3. **Did you notice the METAR status at HGR (Hagerstown)?**
4. What is your plan of action?

Appendix D: Weather Presentation Questionnaire

Weather Presentation Questionnaire

1. Using the weather presentation, how easy was it to see the METAR information?

Very Hard				Very Easy
1	2	3	4	5

2. How easy was it to determine when a METAR symbol changed from VFR to IFR?

Very Hard				Very Easy
1	2	3	4	5

3. To what degree did the METAR information affect your decision to stay on your course or to fly to an alternate destination airport?

Not at All				Very Much
1	2	3	4	5

4. To what degree did the METAR information affect your decision to stay VFR or to request an IFR flight plan?

Not at All				Very Much
1	2	3	4	5

5. How would you rate the benefits of the weather presentation you used to other sources of weather information (ATIS, Flight Watch, etc.)?

Not at All				Very Much
1	2	3	4	5

6. How much do you think the weather presentation decreased your workload?

Not at All				Very Much
1	2	3	4	5

7. How much did you trust the weather presentation to give you correct information?

Not at All				Very Much
1	2	3	4	5

8. How easy was it to determine the position of the aircraft based on the presentation?

Very Hard				Very Easy
1	2	3	4	5

Thank you very much for participating in our study, we appreciate your help.

Appendix E: Practical Trials

Practical Trials
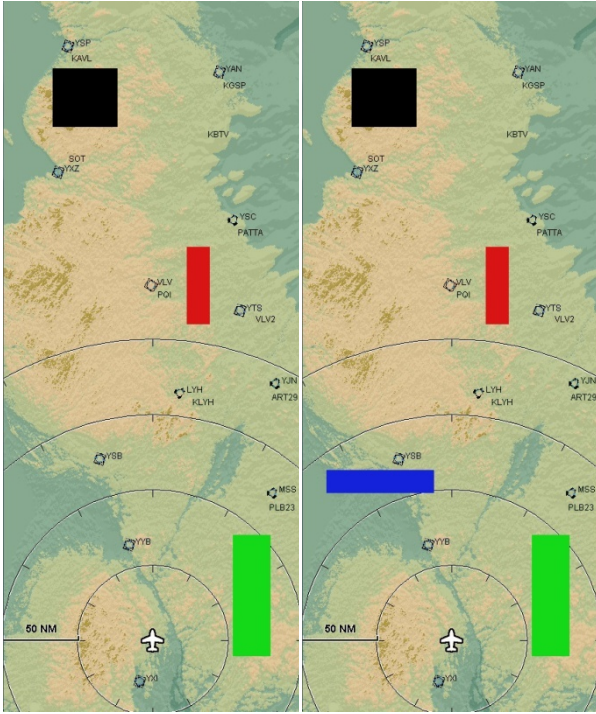
Practice change trial #1 (onset)
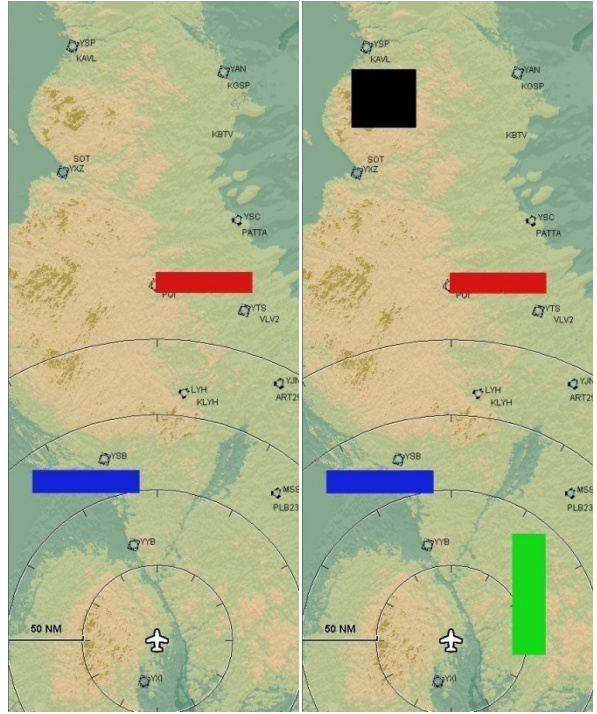
Practice change trial #2 (onset)
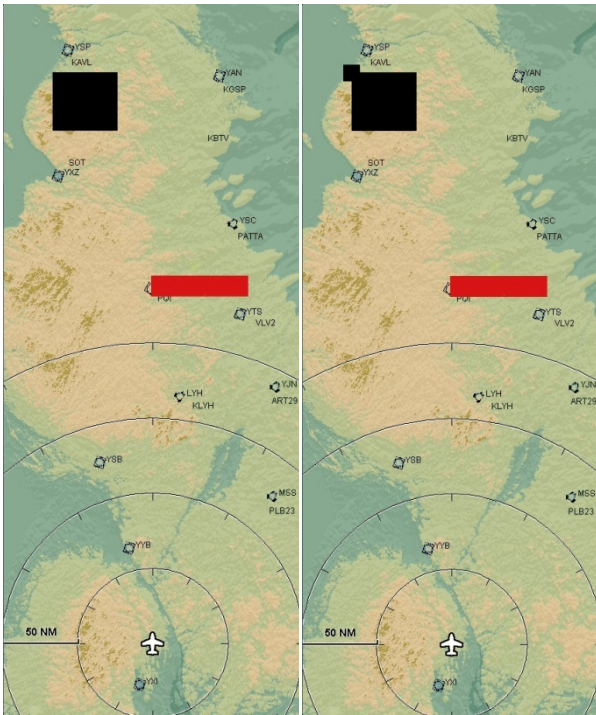
Practice change trial #3 (onset)
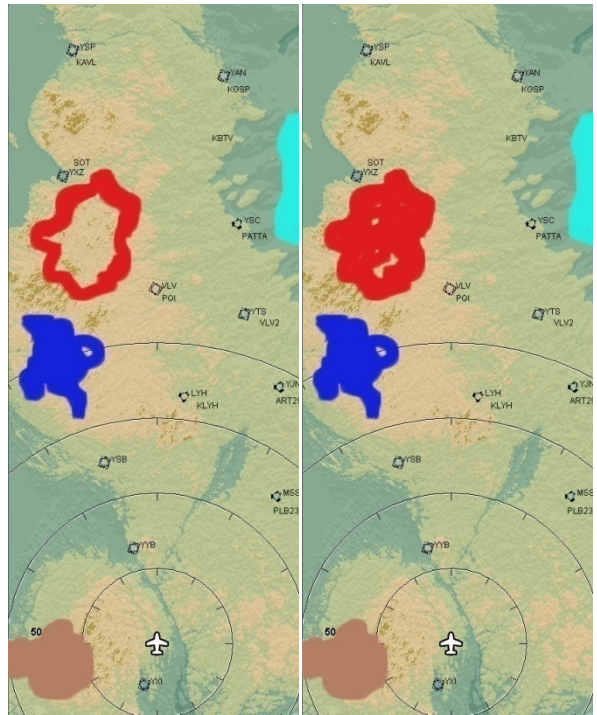
Practice change trial #4 (onset)
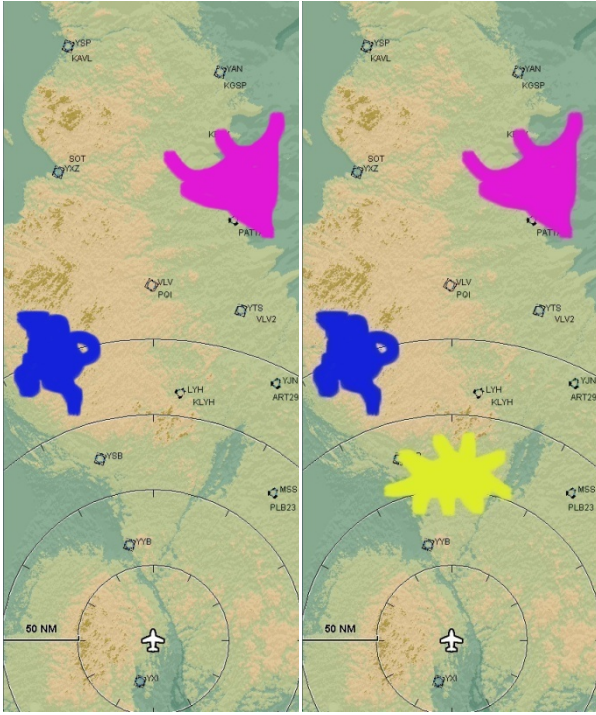
Practice change trial #5 (onset)
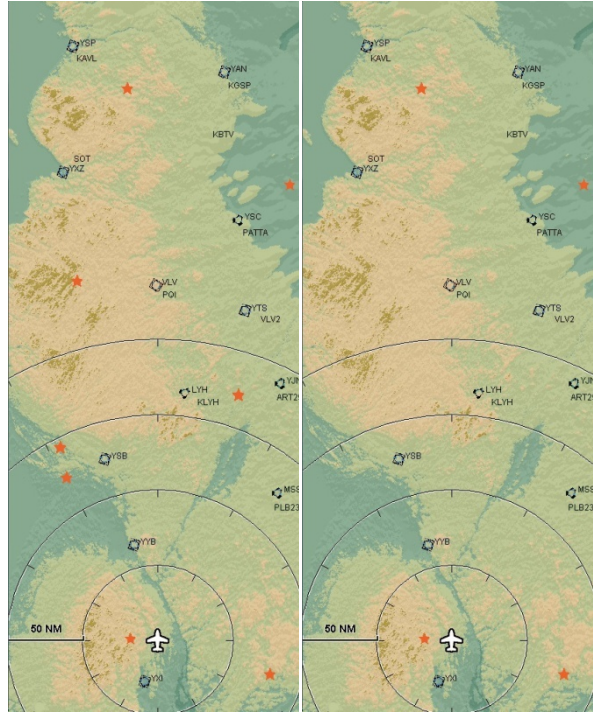
Practice change trial #6 (onset)
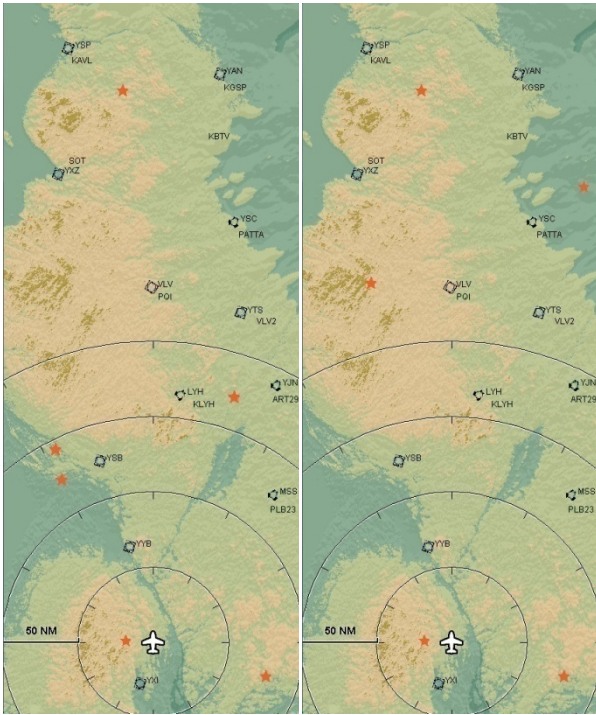
Practice change trial #7 (onset/offset)
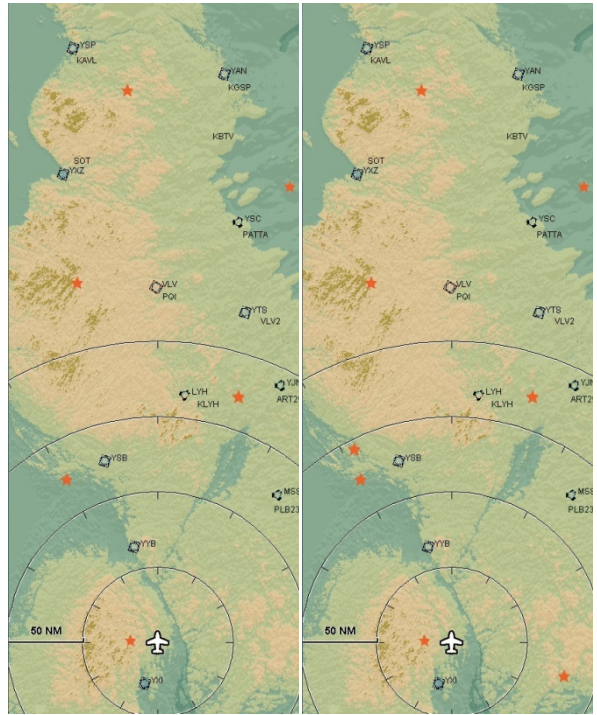
Practice change trial #8 (onset/offset)
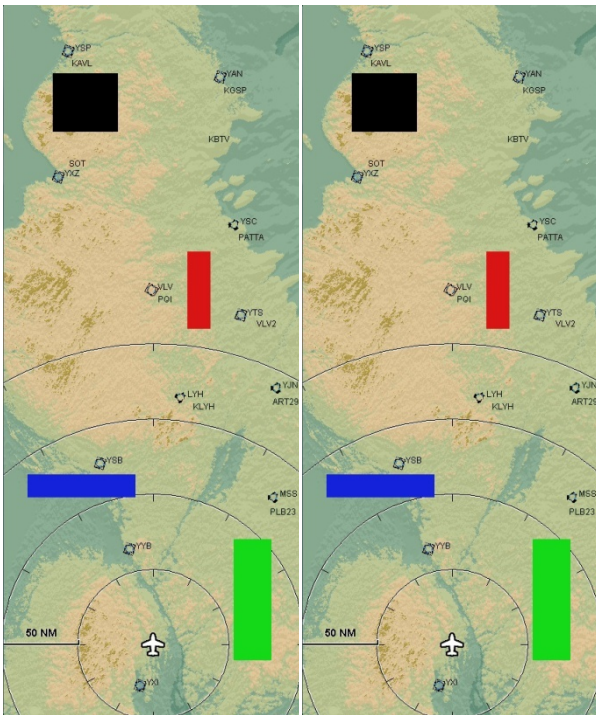
Practice catch trial #1
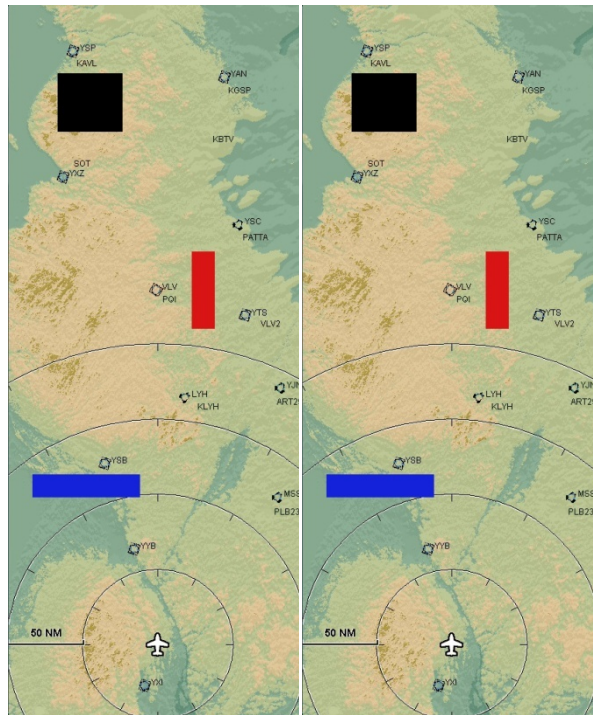
Practice catch trial #2
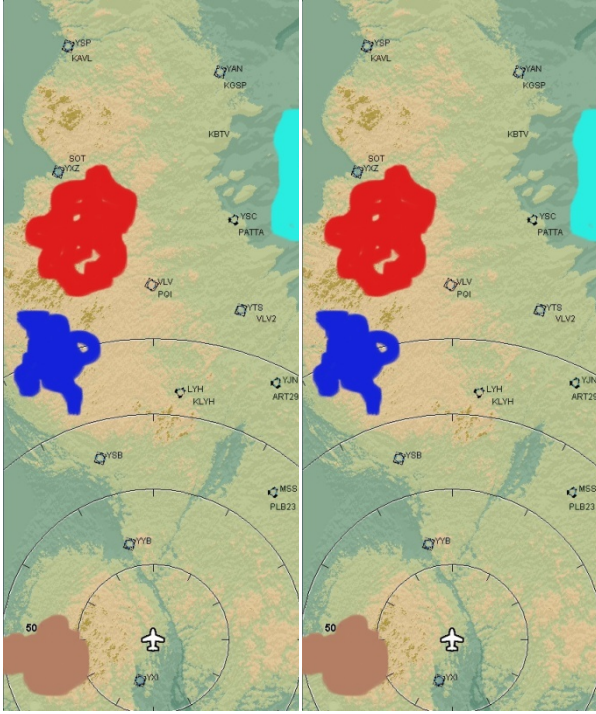
Practice catch trial #3
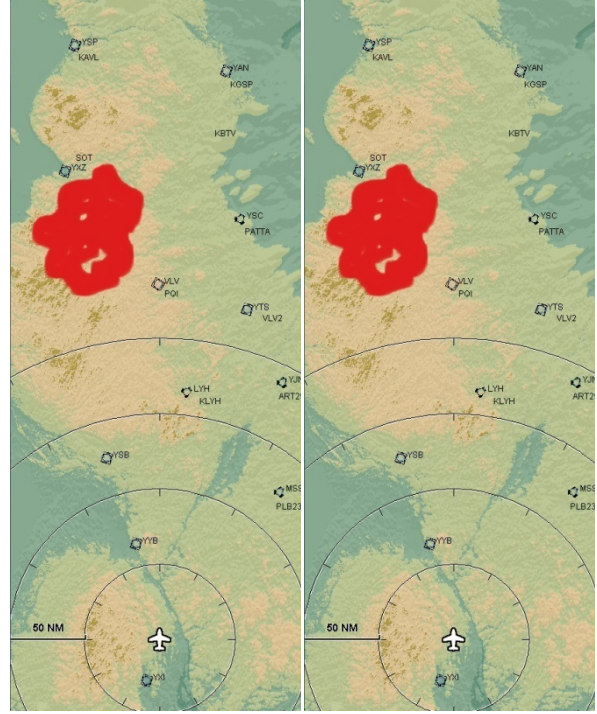
Practice catch trial #4
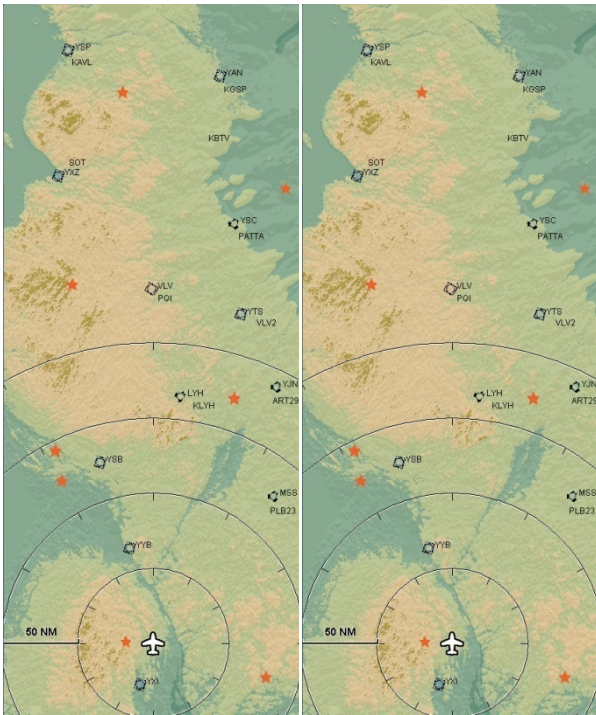
Practice catch trial #5
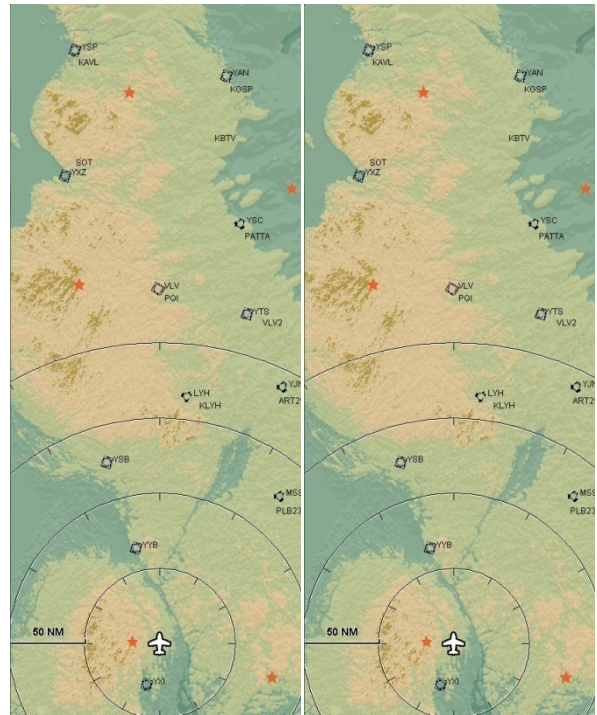

Practice catch trial #6

Appendix F: Experimental Trials
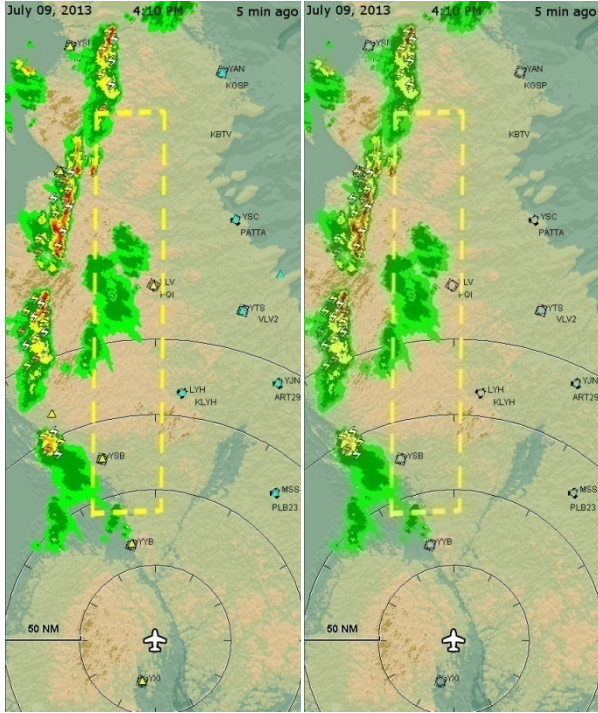
Experimental change trial #1 (WP 1)
(Offset: METARs)
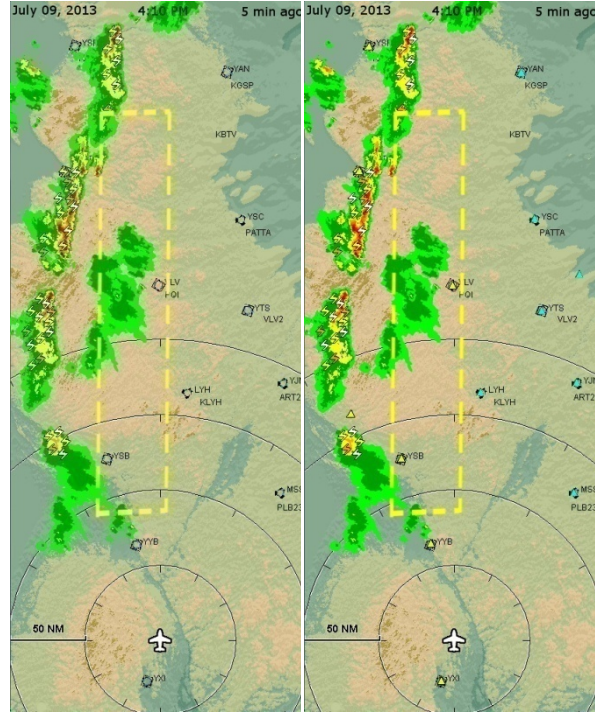
Experimental change trial #2
(Onset: METARs)
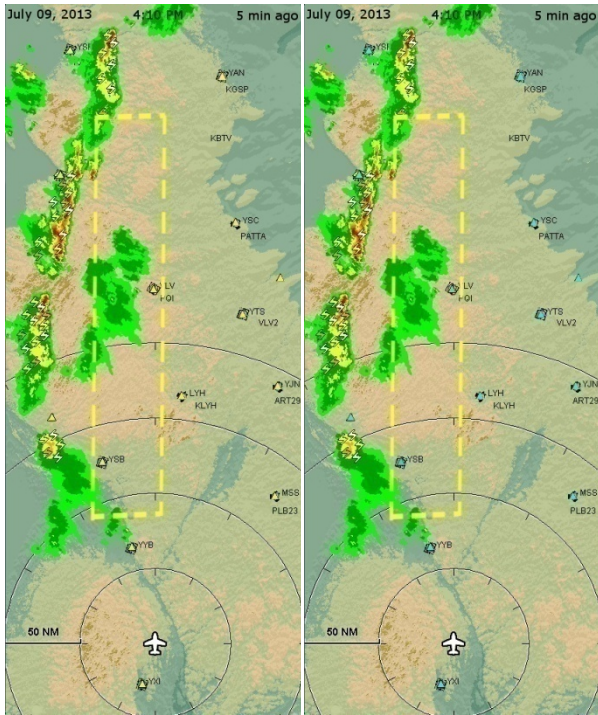
Experimental change trial #3
(Color change: METARs IFR to METARs VFR)
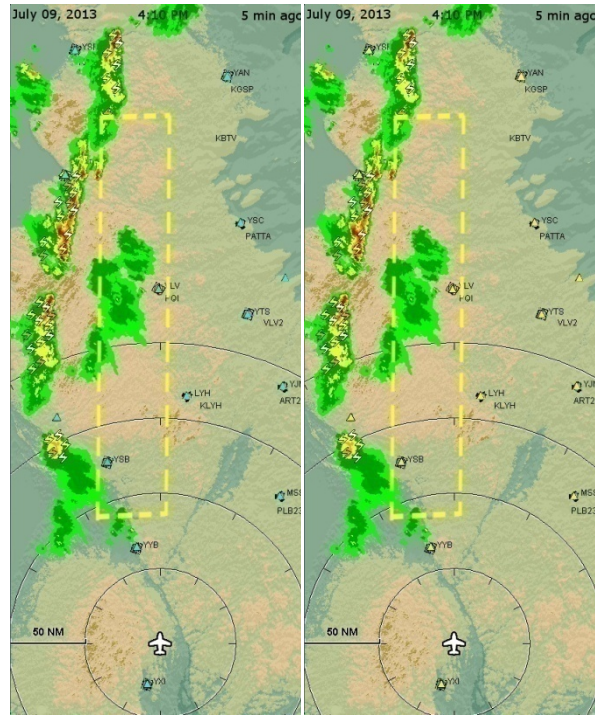
Experimental change trial #4
(Color change: METARs VFR to METARs IFR)
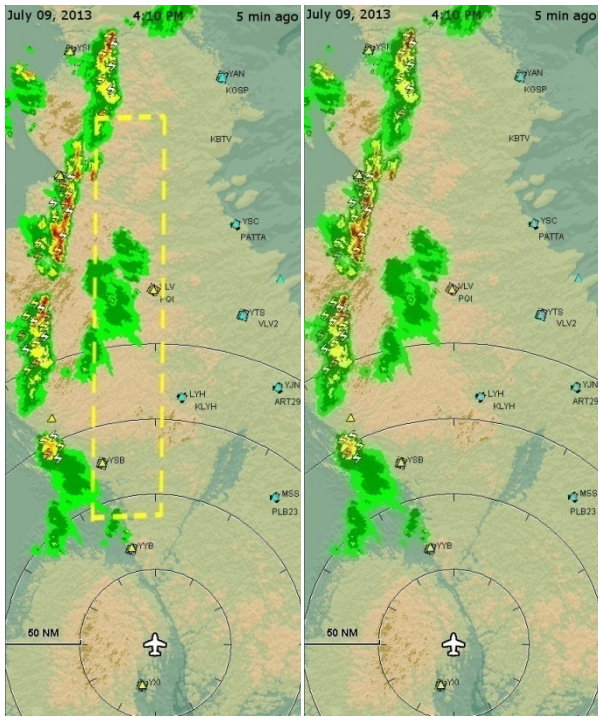
Experimental change trial #5
(Offset: SIGMET)
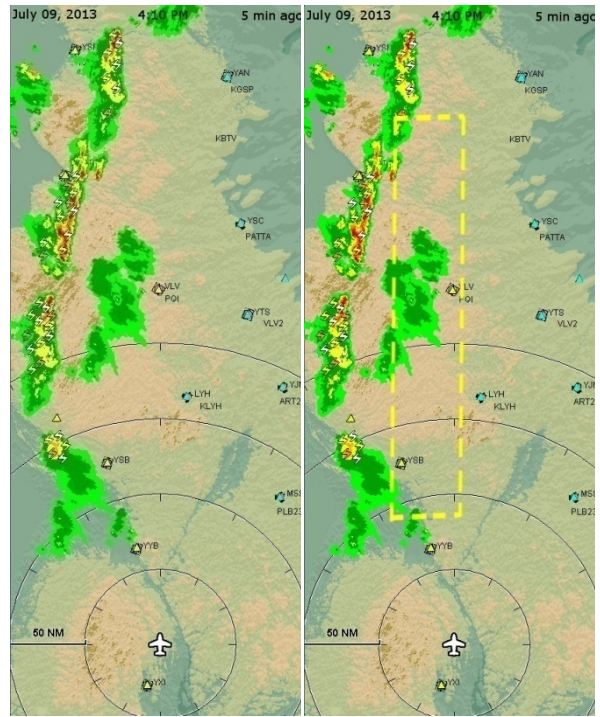
Experimental change trial #6
(Onset: SIGMET)
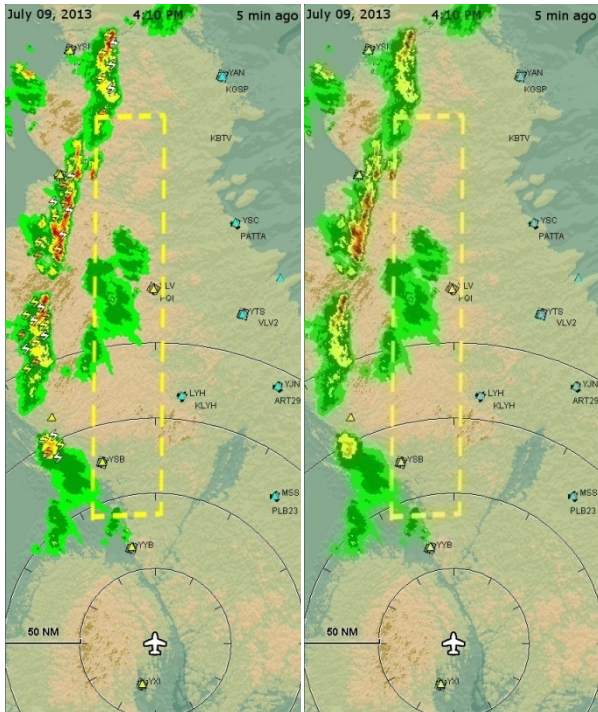
Experimental change trial #7
(Offset: Lightning)
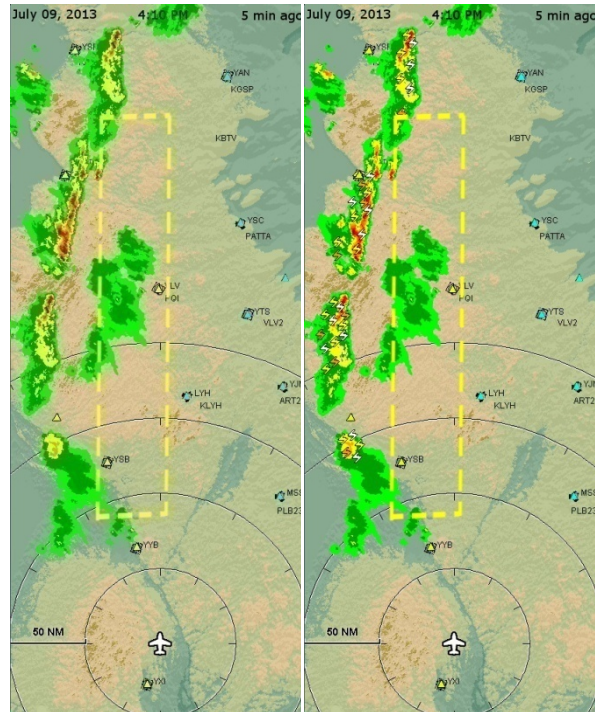
Experimental change trial #8
(Onset: Lightning)
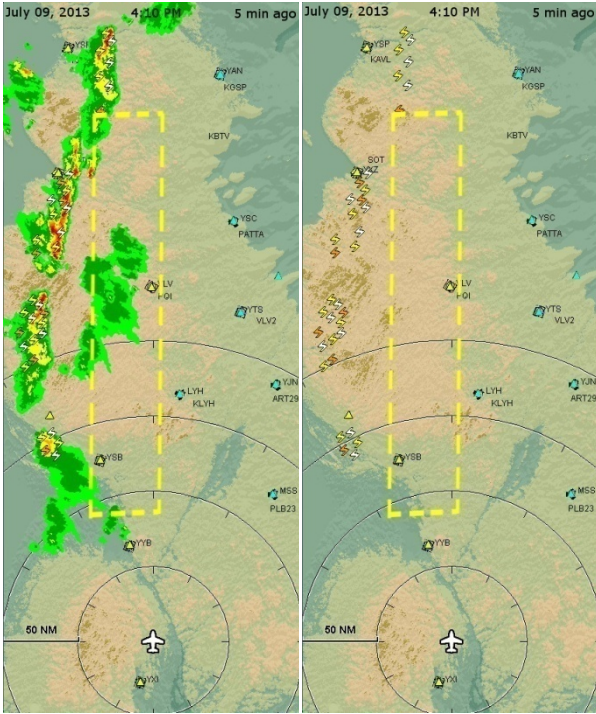
Experimental change trial #9
(Offset: Precipitation)
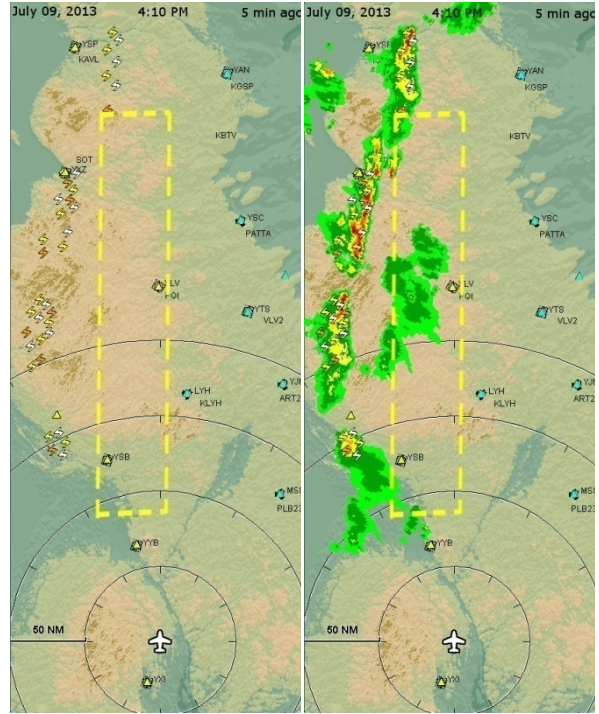
Experimental change trial #10
(Onset: Precipitation)
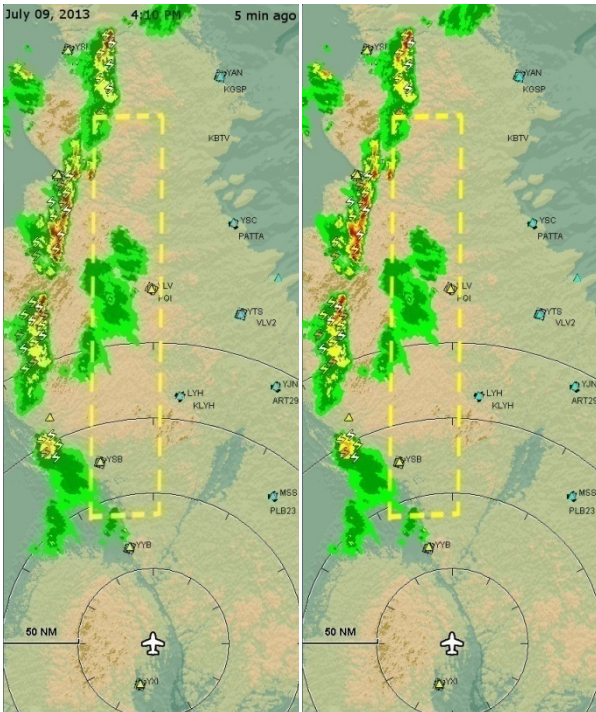
Experimental change trial #11
(Offset: Time-stamp)
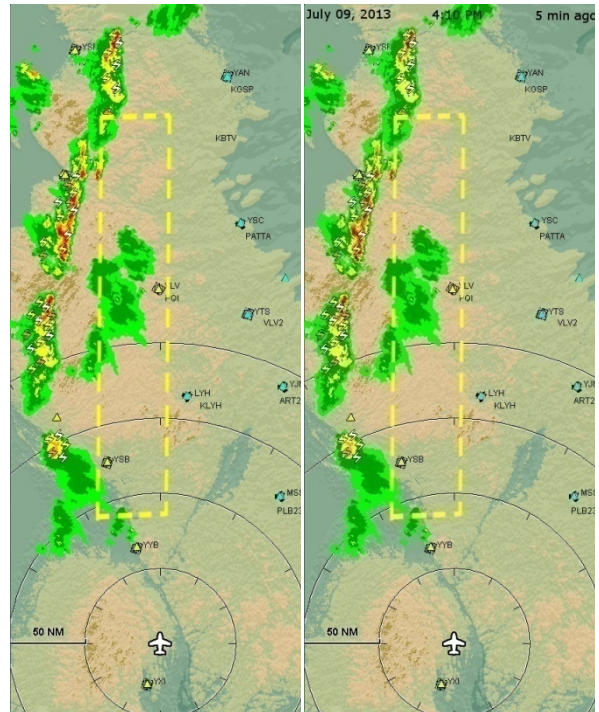
Experimental change trial #12
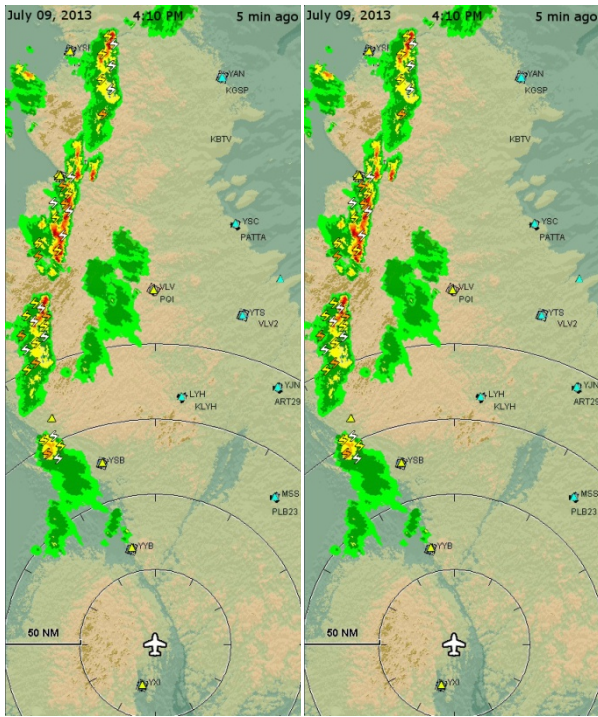(Onset: Time-stamp)
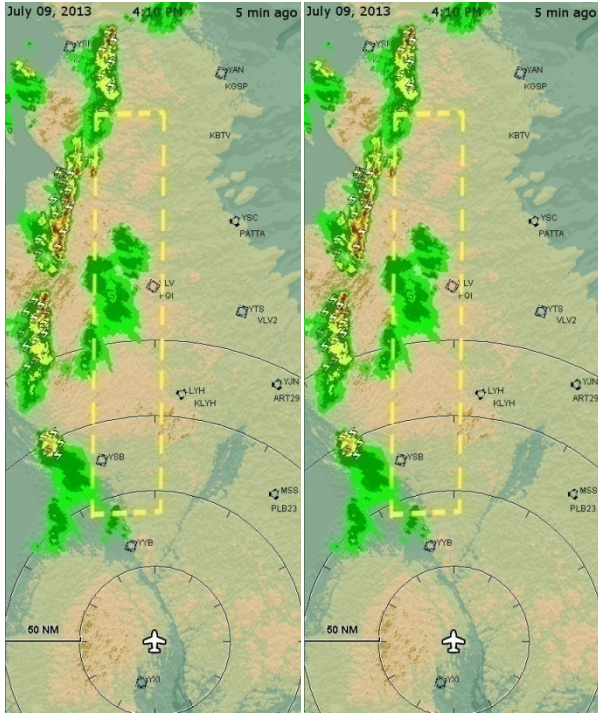
Experimental catch trial #1
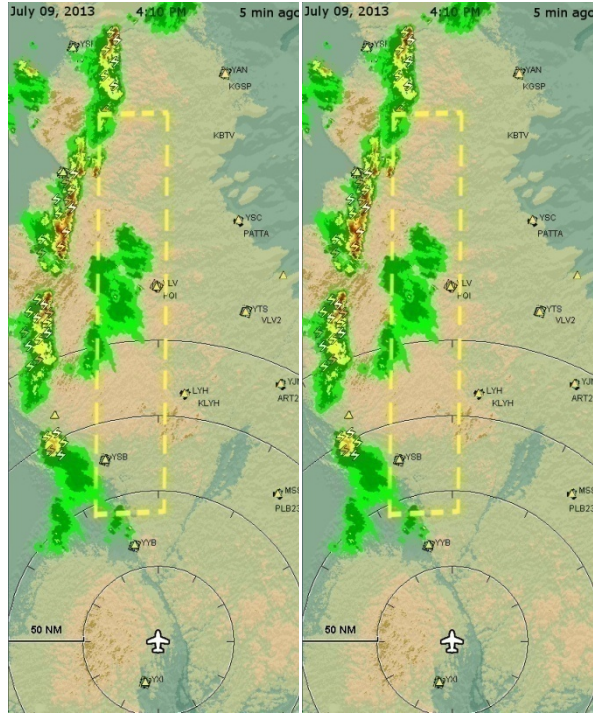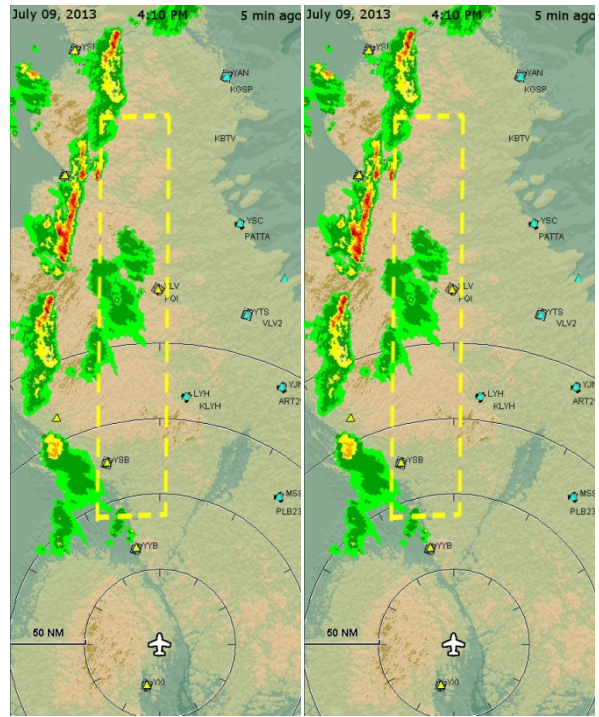

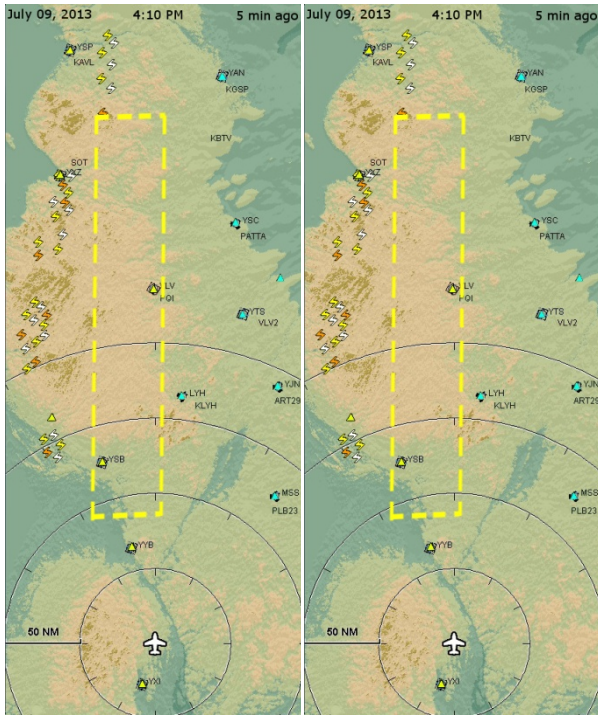Experimental catch trial #2
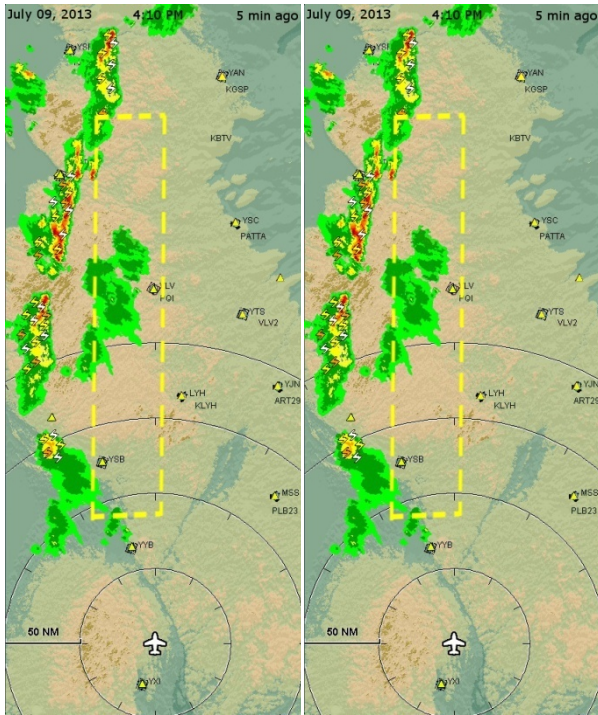
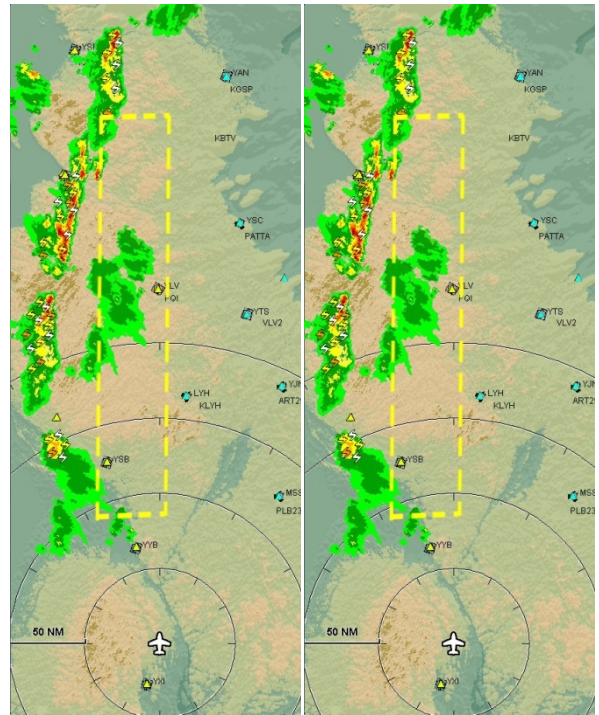Experimental catch trial #3
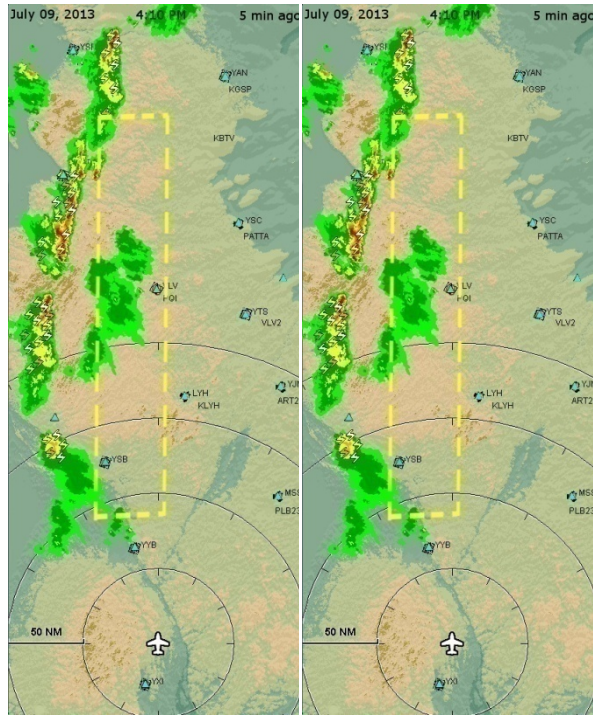
Experimental catch trial #4

Experimental catch trial #5

Experimental catch trial #6

Experimental catch trial #7

Experimental catch trial #8