

**EFFECTS OF PROBABILISTIC FLIGHT-ROUTE RISK  
ESTIMATES FOR ENHANCED DECISIONS (FRREED) ON  
AERONAUTICAL WEATHER-HAZARD DECISION-MAKING**

A Thesis  
Presented to  
The Academic Faculty

by

Sweta Parmar

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science in the  
School of Psychology

Georgia Institute of Technology  
May, 2020

**COPYRIGHT © 2020 BY SWETA PARMAR**

**EFFECTS OF PROBABILISTIC FLIGHT-ROUTE RISK  
ESTIMATES FOR ENHANCED DECISIONS (FRREED) ON  
AERONAUTICAL WEATHER-HAZARD DECISION-MAKING**

Thesis Committee:

Dr. Rick Thomas, Advisor  
School of Psychology  
*Georgia Institute of Technology*

Dr. Bruce Walker  
School of Psychology  
*Georgia Institute of Technology*

Dr. Karen Feigh  
School of Aerospace Engineering  
*Georgia Institute of Technology*

Date Approved: February 26, 2020

## **ACKNOWLEDGEMENTS**

I would like to thank all my lab mates for their continuous support and suggestions during brainstorming and the whole writing process. I would also like to thank my friends and family for listening to my problems and supporting them selflessly.

# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>iii</b>
<b>LIST OF TABLES</b>	<b>vi</b>
<b>LIST OF FIGURES</b>	<b>vii</b>
<b>LIST OF EQUATIONS</b>	<b>viii</b>
<b>SUMMARY</b>	<b>ix</b>
<b>1. INTRODUCTION</b>	<b>1</b>
1.1 Weather-related decision making	2
1.2 Plan Continuation Errors	4
1.3 Weather Products	5
1.3.1 NEXRAD	6
1.3.2 NCWF	7
1.4 Decision making with NEXRAD	8
1.5 Weather-related decision aids in the cockpit	9
1.6 Providing Forecast Uncertainty Information	12
1.7 Probability Rendering of Uncertainty Information using Cumulative Risk	15
1.8 Study Overview	18
<b>2. METHODS</b>	<b>21</b>
2.1 Participants	21
2.2 Materials	22
2.2.1 Training and Test Phase	23
2.2.2 Probabilistic FRREED	27
2.3 Procedure	28
2.4 Theory for Dependent Variables	28
2.4.1 Trust in automation	28
2.4.2 Confidence Judgments and Calibration	30
2.4.3 Brier Score and Brier Skill Score	31
2.4.4 Signal Detection Theory	32
<b>3. RESULTS</b>	<b>34</b>
3.1 Performance (Proportion Correct) Trends across Conditions	34
3.2 Sensitivity and Bias Analysis	35
3.3 Confidence Judgment Calibration	39
3.4 Learning Within-Subject	41
3.4.1 Transfer of Learning	41
3.4.2 Training Phase Learning	42
3.4.3 Test Phase Learning	45
3.5 Trust in Automation	49

<b>4. DISCUSSION</b>	<b>54</b>
<b>4.1 Implications</b>	<b>56</b>
<b>4.2 Limitations</b>	<b>57</b>
<b>4.3 Future Research</b>	<b>58</b>
<b>APPENDIX A. PARTICIPANT INSTRUCTIONS</b>	<b>60</b>
<b>APPENDIX B. PRE-STUDY QUESTIONNAIRE FOR STUDENT PARTICIPANTS</b>	<b>64</b>
<b>APPENDIX C. DEMOGRAPHICS QUESTIONNAIRE FOR PILOTS</b>	<b>65</b>
<b>APPENDIX D. FEEDBACK QUESTIONNAIRE</b>	<b>70</b>
<b>APPENDIX E. NEXRAD STIMULI</b>	<b>71</b>
<b>REFERENCES</b>	<b>72</b>

## LIST OF TABLES

Table 1. Weather-related decision aids .....	11
Table 2. Outcomes table for signal detection theory .....	33
Table 3. Summary statistics for Type 3 generalized estimating equations analysis, displaying flight-path safety decisions regressed on information characteristics and task outcome.....	37
Table 4. Group wise Mean Brier Score and Brier Skill Scores .....	40
Table 5. Summary statistics for Type 3 generalized linear model analysis, displaying Brier Score regressed on information characteristics .....	40
Table 6. Kendall’s Tau and Pearson’s Correlations [ $\tau_b$ ( $r$ )] matrix between training and test phase.....	42
Table 7. Summary statistics for Type 3 generalized linear model analysis, displaying performance in the training phase regressed on information characteristics .....	43
Table 8. Summary statistics for Type 3 generalized linear model analysis, displaying performance in the test phase regressed on information characteristics.....	46
Table 9. Kendall’s Tau and Pearson’s correlations [ $\tau_b$ ( $r$ )] matrix for trust scale items, and Confidence Judgments and Brier Scores.....	52

## LIST OF FIGURES

Figure 1. Example of static NEXRAD .....	6
Figure 2. Example of static NCWF.....	7
Figure 3. Single route trial with moderate-accuracy FRREED .....	23
Figure 4. Elicitation for one double route trial in the training phase (with high accuracy FRREED).....	25
Figure 5. Elicitation for one storm with single-route trials in the test phase (with baseline accuracy FRREED).....	26
Figure 6. Trust in automation scale by Jian et al. (2000).....	30
Figure 7. Effect of level of support on performance.....	35
Figure 8. Trend in mean discriminability across conditions.....	38
Figure 9. Trend in mean bias across conditions.....	38
Figure 10. Trend in mean performance for subsequent training blocks across conditions.....	44
Figure 11. Trend in mean performance for different scenarios across conditions.....	45
Figure 12. Trend in mean performance for subsequent storms presented across conditions .....	47
Figure 13. Trend in mean performance for different scenarios across conditions.....	48
Figure 14. Trend in mean performance for unfolding trials across conditions.....	48
Figure 15. Mean trust in automation score for each condition .....	51

## LIST OF EQUATIONS

Equation 1: Brier Score.....	31
Equation 2: Brier Skill Score .....	32
Equation 3: Sensitivity or Discriminability Index .....	33
Equation 4: Response Bias or Criterion.....	33

## SUMMARY

A tool commonly used to aid the navigational decisions of pilots to avoid weather hazards is Next Generation Radar (NEXRAD), which provides information about geographically referenced precipitation. However, this tool is limited because, when pilots use NEXRAD, they have to infer the uncertainty in the meteorological information for both understanding current hazards as well as extrapolating the impact of future conditions. Recent advancements in meteorology modeling afford the possibility of providing uncertainty information concerning hazardous weather for the current flight. Although probabilistic weather products do not exist in today's cockpit, it is critical to evaluate how operators might use or misuse such products when incorporating uncertainty information in their decision-making. In addition, it is important to study how accurate a probabilistic decision aid needs to be for effective use by operators. Although there are systematic biases that plague professional's use of uncertainty information, there is evidence that presenting forecast uncertainty can improve weather-related decision-making. The current study investigates a simulated probabilistic component of a decision aid that renders flight-path risk as a probability that the route will come within 20 nmi radius (FAA recommended safety distance) of hazardous weather within the next 45 minutes of flight. The study evaluates four NEXRAD displays integrated with Flight-Route Risk Estimates for Enhanced Decisions (FRREED) providing varying levels of support. The "no" support condition has no FRREED (the NEXRAD only condition). The "baseline" support condition employs a FRREED whose accuracy is consistent with current capability in meteorological modeling. The "moderate" support condition employs a FRREED whose

accuracy is likely at the top of what is achievable in meteorology in the near future. The “high” support display provides a level of support that is likely unachievable in an aviation weather decision-making context without significant technological innovation. The results indicate that operators did rely on the FRREED to improve their performance over the no-support condition (NEXRAD only). The level of performance of the operators improved in terms of both calibration and resolution as the aids increased in accuracy. I will discuss the implications of the findings for the safe introduction of probabilistic decision aids in future general aviation cockpits.

# 1. INTRODUCTION

General aviation comprises of all civilian aircraft except the scheduled passenger aircraft. An extraordinarily high number of deaths occur in general aviation accidents. The accident rate in general aviation is very high: 6.51 per 100,000 flight hours (NTSB, 2014). In 2011, 18% of the general aviation accidents were fatal for 444 people (NTSB, 2014). General aviation accidents that year led to monetary losses between 1.64-4.64 \$ billion (Sobieralski, 2013).

The National Transportation Safety Board (NTSB) (NTSB, 2014) identified that one of the major factors in general aviation accidents is hazardous weather. The Federal Aviation Administration (FAA) has found that 20% (8657) of the general aviation accidents from 2003 to 2007 had weather as one of the major contributing factors (FAA, 2010). Weather is an integral part of pilots' flight operation, even with in-cockpit weather radar, it continues to pose a threat to flight operations and general aviation safety. This threat can be attributed both to the pilot and system errors.

An analysis of the NTSB accident database from 1994-2014 on thunderstorm-related accidents revealed that convective weather is involved in 70% of fatalities (Boyd, 2017). The common factors amongst these fatalities identified by Boyd are non-adherence of pilots to FAA recommended minimum-separation-distance of 20 nmi from hazardous weather and tendency of pilots to select a riskier route (downwind) for circumnavigating convective weather. These factors highlight a need to identify the root cause behind pilot failures and provide proper aids accordingly. A considerable number of decision support tools have already been designed to decrease fatalities, but they still fail to reduce accidents. One of the reasons behind continuing failures in enhancing pilots' decision-making is that most of the decision aids are deterministic because they leave

behind the uncertainty in the meteorological weather data for the pilot to interpret. A solution to this issue is to provide uncertainty information to pilots to increase their situation awareness about presented meteorological data. This thesis evaluates a novel probability-based Flight-Route Risk Estimates for Enhanced Decisions (FRREED) that provides a quantitative estimate of the uncertainty in meteorological weather data.

The current study investigates a probability-based FRREED simulated for the NEXRAD display that can make pilots aware of the uncertainty in the weather data. The NEXRAD is a radar display, which provides geographically referenced precipitation activity. The FRREED provides the probability of a particular flight route intersecting the hazardous weather at least once (cumulative risk) during its course of flying. Hence, supports the user to select a safe route.

Firstly, this thesis discusses the complexity of weather-related decision-making, and then it mentions how decision-making is prone to certain kinds of errors. Following this, there is a discussion on currently available weather products and their limitations. This discussion helps in establishing the importance of decision aids in general aviation cockpits with special attention to the need for probabilistic decision aids to enhance the decision-making of pilots in future cockpits. Lastly, the document discusses the description of the current study, methods used, results obtained, and the conclusion drawn.

## **1.1 Weather-related decision making**

Weather-related decision making is a complicated process. Operations in inclement weather require pilots to make multiple evaluations and decisions including identification of the presence of a hazard, estimation of the proximity of weather, estimation of impact to the flight path and taking of appropriate actions. The weather-related information that the pilot receives generally

arrives from a variety of sources that include meteorological briefings, inflight weather reports, visual information from the cockpit, and the on-site reports (Hunter, Martinussen, & Wiggins, 2003). Because of the dynamic environment of aircraft cockpits, these data change very quickly. Pilots have to monitor for changes and update their mental representation of the situation accordingly to maintain situational awareness.

In the last few decades, there have been significant advances in improving designs for weather displays in both general and commercial aviation. However, weather displays are a complex system that comprises multiple weather information sources that have different spatial and temporal resolutions. There are multitude of problems that exist with current weather information systems like lack of remote sensing capability, lack of pilot reports, imprecise diagnostic products, imprecise forecasts, etc. (Bass & Minsk, 2001). This complexity makes it difficult for pilots to make swift and sound decisions during hazardous weather conditions (Elgin & Thomas, 2004). The situation is worse for general aviation because these aircraft are mostly smaller planes flying at low altitudes with low speeds, which make them more prone to adverse weather conditions (Wu, Ligda, Lim, Comerford, & Johnson, 2010).

The main focus of this thesis is general aviation weather displays. The most common cause behind weather-radar-related accidents is either a failure in recognizing or in interpreting the weather information displayed in the cockpit (Wiggins, 2005). Even a small change in weather display symbology (colors and symbols) among general aviation pilots leads to credible differences in weather deviations, cognitive engagement, and weather display usage (Ahlstrom, 2015). This evidence shows that a pilot's understanding of weather data is fluid, and it can be affected by multiple factors. The first step towards solving the complexity involved in the weather-related decisions is understanding errors in decision-making and common causes behind those

errors. The next section elaborates on plan continuation errors, which is the most common decision-making error that pilots make.

## **1.2 Plan Continuation Errors**

Research shows that even after pilots receive information that signals a need to revise their plan, they often continue to follow the original flight plan (Orasanu, Ames, Martin, & Davison, 2001). This tendency is known as plan continuation errors. While following a flight plan, pilots should still monitor the airspace continuously to look for changes in airspace that might cause them to revise the initial flight plan. Plan continuation errors are more common with pilots that inadequately monitor their airspace. Muthard and Wickens (2002) found that pilots in simulations failed to revise their flight plan in response to hazardous elements in the airspace approximately one-third of the time and that pilots who demonstrated inadequate monitoring were most likely to commit these errors of omission.

Plan continuation errors have also been observed when pilots decide on whether to continue from Visual Flight Rules (VFR) to Instrument Meteorological Conditions (IMC) or to divert from their path (Wiegmann, Goh, & O'Hare, 2002). The same study showed that pilots who observed deteriorating weather early in their flight path were more optimistic about the weather and chose to continue through it for an extended amount of time before deciding to divert. This delayed decision to divert was also influenced by their previous flying experience, with more experienced pilots deciding to divert earlier than less experienced pilots. This study showed that pilots are prone to underestimating the risk of continuing in deteriorating weather and that their understanding of risk depends on their expertise. These findings indicate a need to provide pilots with proper risk

information about flight paths and a need to train pilots so that they can make better flight path decisions in deteriorating weather.

The evidence mentioned in this section shows that plan continuation errors are quite common in aviation. Some of the major underlying causes behind these errors are inadequate plan monitoring, underestimation of risk, inexperience, and so on. This thesis addresses the issue of plan monitoring and risk estimation by simulating and evaluating a probabilistic component of a decision support tool that, when used with NEXRAD, helps reduce plan continuation errors. Before describing this component, in the next section, I will elaborate on some of the issues in the current cockpit weather technology.

### **1.3 Weather Products**

As discussed previously, plan continuation errors are known to affect decision making, performance, and safety of flight operation. Dutcher and Doiron (2008) emphasized the need to understand the reasons why pilots fail to take appropriate actions during deteriorated weather. They identified some of the possible causes were the pilots' lack of knowledge about meteorological products and meteorology, limitations of technology, and inadequate pilot training to deal with the weather.

A common theme in the literature is that pilots require meteorological products that provide not only real-time weather information but also adequate decision-making support as they multitask in a dynamic environment. However, this requires a proper representation of weather (e.g., graphical representation) with high spatial accuracy and temporal resolution. The following subsections will discuss some common weather products (with the emphasis on NEXRAD), how they support decision-making at varying levels, as well as identify their limitations.

### 1.3.1 NEXRAD

One of the most used displays in general aviation cockpits is NEXRAD. Fig. 1 shows a sample NEXRAD image (NOAA). It provides geographically referenced precipitation activity uplinked to cockpits through Flight Information Services Data Link (FISDL) (Wu et al., 2010). However, one of the major issues with NEXRAD is its latency in that the data reaches the cockpit 5-7 minutes after it is received from the service provider and is refreshed in the cockpit only once every 5-7 minutes (Elgin & Thomas, 2004). Pilots have to use this 14-minute-old data to extrapolate current weather conditions and predict future conditions. Typically, pilots do not adequately account for this temporal delay, and they treat the information like real-time, which hampers their decision-making process. Another drawback of NEXRAD is that it does not help pilots understand the effect of weather cell movements on the flight path, which pilots have to estimate on their own. Thus, increasing the chance that during bad weather, pilots will overestimate or underestimate nearby weather risks, which can lead to plan continuation errors.

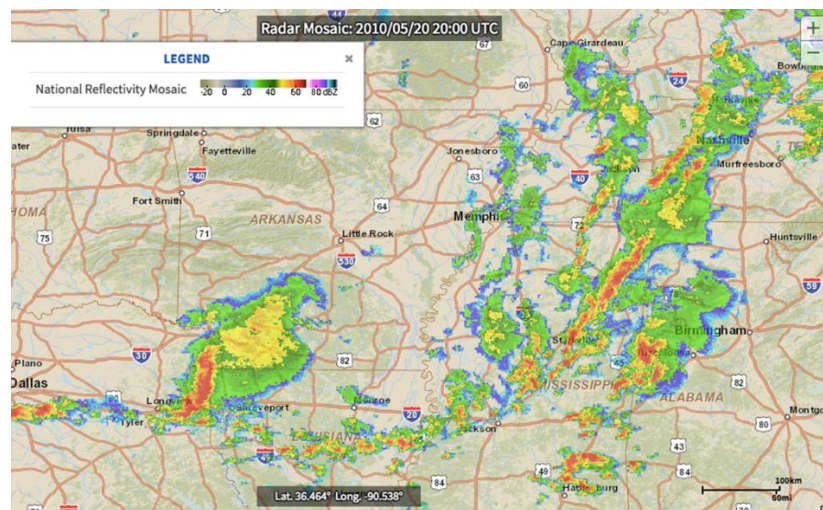


Figure 1. Example of static NEXRAD

### 1.3.2 NCWF

National Convective Weather Forecast (NCWF), developed by the National Center for Atmospheric Research, provides current weather hazard information (using NEXRAD) and one-hour prediction of hazardous weather through extrapolated polygons, as shown in Fig. 2. NCWF focuses only on large areas of convective activity, not on areas less than 512 km<sup>2</sup>. The smoothed forecast contours have an approximate spatial resolution of 8 km (NCWF, 2007), which does not allow forecast for smaller convective areas that can be of importance in hazardous airspace. In this way, NCWF is only a deterministic forecast product in that it makes the pilots understand or interpret levels of uncertainty on their own. This thesis provides a solution to this limitation of NCWF by proposing a novel probabilistic Flight-Route Risk Estimates for Enhanced Decisions (FRREED) that accounts for uncertainty in forecasting.

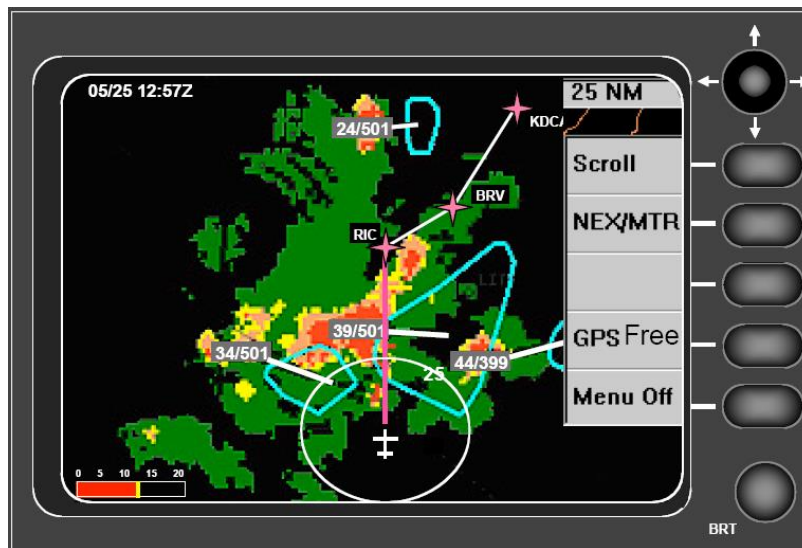


Figure 2. Example of static NCWF

## 1.4 Decision making with NEXRAD

As discussed earlier, NEXRAD provides precipitation information near the aircraft to support pilots with decision making. NEXRAD provide important perceptual information to help pilots deal with hazardous weather conditions, especially the location of hazardous weather cells around their flight path. As mentioned previously, the biggest challenge with NEXRAD, which has been studied extensively in the past, is the latency in the data that it. However, Knecht (2016) reviewed multiple NEXRAD studies and identified another major challenge that when using NEXRAD, pilots overestimate their distance from storms and underestimate the risk, which eventually leads them dangerously close to the hazardous weather. However, to date, the literature has not emphasized this underestimation of risk or provided satisfactory ways to make pilots aware of the uncertainty in forecasts. This risk estimation is further degraded because pilots have to manually estimate their distance from the storm. Knecht (2016) recommended a need for a 20-nmi range ring along with the own ship icon for the NEXRAD display.

Many studies have looked at the limitations of NEXRAD. Yuchnovicz, Novacek, Burgess, and Stokes (2001) compared the use of NEXRAD with conventional pre-flight and in-flight weather products and found that, when compared to conventional products, NEXRAD does not improve pilot's decision making. They concluded that even though NEXRAD provides pilots with more information, this additional information increased the pilots' mental workload and decreased their time to make decisions. Beringer and Ball (2004) found that pilots with high-resolution NEXRAD images spent more time looking at them compared to pilots who had only low-resolution images. They found that the pilots with high-resolution images were also more likely to continue to fly in deteriorated weather as they believed that they could fly around or in between ominous weather cells. In the same study, 53% of pilots failed to comply with the official FAA

guideline (FAA, 2013) that requires pilots to maintain at least 20-nautical mile (nmi) separation from hazardous weather.

Burgess and Thomas (2004) demonstrated the inadequacy of current weather displays in aiding pilots during hazardous weather by providing pilots with improved weather displays equipped with either the NEXRAD or the NCWF. Weather-avoidance behavior of pilots was not significantly different between the two display conditions. They found that all the groups maintained a mean distance of 10 nmi from hazardous weather cells, which is half of the FAA recommended safety margin. They suggested that proper pilot training and display design improvement (in terms of proper representation of necessary information) are required steps in enhancing pilot decision-making and pilot performance. According to Burgess and Thomas (2004), the current weather displays are not designed specifically to support pilots' higher level decision making because these displays do not help pilots to understand the impact of hazardous weather on their flight path. They determined that the displays provide pilots only with cues to predict the impact of hazardous weather by themselves. Hence, there is a need for decision aids in cockpits that can support both higher-level decisions and subsequent actions. The next section mentions some of the currently available decision aids, and their limitations and is followed by a description of a critical component that could potentially reduce some of the limitations of current decision aids, particularly limitations supporting spatial-temporal uncertainty.

## **1.5 Weather-related decision aids in the cockpit**

The literature on decision support systems has a wide range of systems, from small-scale interventions to large-scale technical systems. Small scale interventions often provide extra alerts or notification to pilots through portable devices like iPads and Android phones. Large-scale

systems include weather displays with advance alerting information or decision-aiding cues to guide pilots' decision-making process. Table 1 discusses some of the small- and large-scale decision aids designed for cockpits. Table 1 shows that most of the current small-scale aids inadequately improve pilot's decision making, provide insufficient (or zero) knowledge about the uncertainty associated with a forecast, and do not help pilots follow the FAA guideline of 20 nmi. It shows that some common large-scale decision support systems in general aviation cockpits have also been found to provide inadequate support to pilots' decisions. Thus, signals a need for improved information representation for current and future trajectories of aircraft to support pilots' decision.

Table 1 shows some of the large-scale aids that have attempted to implement probabilistic weather forecasts. However, these decision aids are inadequate in supporting the pilot's decision making. Although these aids emphasized the need for probabilistic weather and uncertainty information, none of them provide any quantitative estimate of risk or uncertainty. Apart from these aids, not much has been done to develop probabilistic aids for pilots because probabilistic aids are difficult both to design and to implement. Calculating probabilities for weather data and presenting them in a comprehensible way to prevent underestimation or overestimation of risk is a complicated task that requires complex ensemble modeling. This thesis approaches this issue by simulating a predictive decision support tool that provides uncertainty information about weather-cell movements to pilots. The next few sections discuss the concept of uncertainty and cumulative risk, along with the reasons why pilots fail to understand this information, what form of representation could make pilots more aware of this risk, and how this representation can be implemented and improved through our proposed predictive tool.

Table 1. Weather-related decision aids

Study	Weather-Related Decision Aids	Limitations
<b><u>Small-Scale Decision Aids</u></b>		
Campbell and Alexander (2016)	In a cockpit simulator-based task, notification regarding nearby weather hazard leads to decrease in response time for identifying hazards.	<ul style="list-style-type: none"> <li>• Although the notifications facilitated response time to hazardous weather, the decision aid did not support pilots to follow the FAA recommended safety margin.</li> </ul>
Ahlstrom, Ohneiser, and Caddigan (2016)	In a simulated flying task, pilots had a portable weather application that leads to increased weather situation awareness (WSA) and cognitive engagement (prefrontal oxygenation levels).	<ul style="list-style-type: none"> <li>• Pilots traveled a bit farther from the hazardous weather, but still violated the FAA guideline</li> <li>• No information about uncertainty was available to pilots.</li> </ul>
<b><u>Large-Scale Decision Aids</u></b>		
Alam, Abbass, and Barlow (2006)	The ant-colony optimization algorithm was used to find hazardous-weather free routes in the airspace.	<ul style="list-style-type: none"> <li>• Multiple versions of the algorithm developed by various researchers can lead to different results.</li> <li>• No information about inherent uncertainty in prediction.</li> </ul>
Spirkovska and Lodha (2002)	Aviation Weather Data Visualization Environment (AWE) system provides graphical weather data in pilot's region of interest to make go or no-go decision on the current path.	<ul style="list-style-type: none"> <li>• The system was specific to the pilot's region of interest and can fail to provide the information regarding the overall generic weather scenario.</li> <li>• Can lead to underestimation of risk for the regions that pilots do not focus on.</li> </ul>
Kronfeld (2003)	Rockwell Collins Enhanced Weather Radar (EWxR) system combines NEXRAD and onboard radar to overcome the attenuation and range limitations of the onboard radar. It also characterizes the cells as hazardous, possibly-hazardous, or non-hazardous based upon attributes like reflectivity level, storm speed, and height.	<ul style="list-style-type: none"> <li>• The system indicates uncertainty by stating weather cells as "possibly-hazardous." However, the extent of uncertainty in quantitative format is missing.</li> <li>• Understanding uncertainty depends on comprehension of the word "possibly-hazardous", leading to underestimation or overestimation of risk.</li> </ul>
Busquets, Ruokangas,	Rockwell Collins' AWARE (Aviation Weather Awareness and Reporting	<ul style="list-style-type: none"> <li>• It only notifies the pilot about the type of hazard and location on the</li> </ul>

and Kelly III (2005)	Enhancements) system automatically analyzes the hazards in the weather data to alert pilots of impending hazards (like precipitation, icing, and turbulence) on their path.	map but does not provide any information about whether this hazard region is low/high-risk region.
Matthews and DeLaura (2010)	Convective Weather Avoidance Model (CWAM) developed to provide deterministic and probabilistic weather avoidance fields (WAFs).	<ul style="list-style-type: none"> <li>• Even after getting both deterministic and probabilistic information, pilots were found to underestimate the risk associated with WAFs and intersect WAFs to avoid factors such as delay of the flight.</li> </ul>

**1.6 Providing Forecast Uncertainty Information**

Pilots often have trouble understanding the amount of uncertainty in many meteorological weather products because these products typically contain no explicit statement of the inherent uncertainty in the forecast (Gigerenzer, Hertwig, Van Den Broek, Fasolo, & Katsikopoulos, 2005). Typically, weather forecasts for aviation provide precision through the use of numerical forecast products, which are “deterministic” in that they provide 100% probability forecasts of an adverse weather event at a specific time in the future. For instance, NCWF is a numerical forecast product that provides a 100% probability of a weather event through extrapolated polygons. This characteristic makes NCWF, and other numerical forecasts leave most of the uncertainty in weather prediction to be inferred by the pilot. Unfortunately, past demonstrations have indicated that pilots tend to interpret weather information as nearly certain information (Burgess & Thomas, 2004; Yuchnovicz et al., 2001).

Advanced weather products do not provide information regarding weather forecast uncertainty in a clear and intuitive manner. Nor do they provide decision support tools that can help pilots adapt their actions to the uncertainty inherent in meteorological products. The

underlying principle in the current study is that weather-related decision-making and performance is inextricably linked to how well pilots understand the uncertainties in weather data.

Recent probabilistic weather products have tried to provide uncertainty information directly to pilots through cockpit information displays. Improvements in this capability will help pilots better understand that weather forecasts are not absolute, which will consequently improve aviation safety and efficiency. Probabilistic forecasts will enable better calibration of people's decisions regarding weather uncertainties, which may significantly reduce pilot workload and improve both flight execution and National Aviation Services (NAS) operations. However, this uncertainty information must be introduced into the cockpit carefully as behavioral decision theory has documented numerous systematic biases in the way humans process uncertainty information (Gilovich, Griffin, & Kahneman, 2002). In the context of air traffic management, Gibbons, Holst, and Jonsson (2013) documented that although expert NAS operators understand that forecasts have errors and are uncertain, the operators are biased in their interpretations of weather information uncertainty. In other words, even when the operators in the Gibbons et al. study were aware of weather uncertainty; they were unable to include that in decision making.

Not much has been done in the aviation context to understand the impact of weather uncertainty information on decision-making. However, some studies have demonstrated that forecast uncertainty information has been found to increase the understanding of weather forecasts, which leads to better calibrated and more beneficial decisions for both expert meteorologists and lay people (Joslyn & LeClerc, 2012; Joslyn, Nadav-Greenberg, Taing, & Nichols, 2009; Joslyn & Nichols, 2009; LeClerc & Joslyn, 2015; Nadav-Greenberg & Joslyn, 2009). Nadav-Greenberg and Joslyn (2009) demonstrated that repeated exposure to weather uncertainty information via training followed by feedback helps people learn to use the available uncertainty information for their

decisions. This finding, which is the major motivation for this thesis, led us to manipulate the interval of feedback in our training and test phase to understand how uncertainty information is utilized differently in each phase.

Joslyn and Nichols (2009) argue against the long-believed notion that lay people find it easier to understand a frequency format compared to a probability format. They argue that all the studies favoring frequency formats ask participants to estimate the level of uncertainty on their own. However, daily weather forecast information provides uncertainty information and expects people to incorporate that information into their decision-making process. Joslyn and Nichols (2009) tested the effects of providing uncertainty information in both probability (90% chance of high wind speed) and frequency (9 out of 10) formats for the wind speed forecast of the day. They found that, contrary to previous research, the information presented in frequency formats was the most difficult for people to utilize to understand the wind speed warnings. Frequency formats led to the largest number of errors in both people's decision making and their understanding of uncertainty. They observed that with the probabilistic format, people were better able to understand wind speed forecasts and make better decisions. These findings provide a good base for the chosen probabilistic format of the uncertainty-based FRREED proposed in this thesis.

The studies mentioned so far in this section emphasize the need for uncertainty information in weather forecasts. This requires a quantitative measure of uncertainty. In general, risk can be defined as the "combination of the probability of a consequence and its magnitude" and uncertainty is defined as the "lack of knowledge concerning outcomes that may result from an imprecise knowledge of the risk" (Willows, Reynard, Meadowcroft, & Connell, 2003). Interpreting uncertainty about events and associated outcomes requires risk assessment (Willows et al., 2003). It is especially needed when the future of storms is quite uncertain and the consequences could be

life threatening. Risk assessment is possible from data available from past events and can be used to calculate the risk and uncertainty of a future event (Willows et al., 2003). The probabilistic FREED simulated in this thesis will provide risk assessment regarding the flight route to make pilots aware of the risk and uncertainty related to a weather event. The FREED is intended to make operators aware of the cumulative risk associated with their exposure to hazardous weather throughout their flight.

### **1.7 Probability Rendering of Uncertainty Information using Cumulative Risk**

An important characteristic of risk is that it is cumulative whenever there is either continuous exposure to the same risk factors for a long duration (e.g., smoking for years) or simultaneous exposure to multiple factors (e.g., smoking and drinking). However, people often perceive these risk factors as independent of each other, which leads to the underestimation of risk. Unfortunately, no studies have been done on cumulative risk in weather displays or aviation in general. In this section, I will cite some of the important findings from some other domains like healthcare, natural hazards, and so on to demonstrate the underestimation of cumulative risk.

Quantitative risk-communication is a critical component of medical decision-making because informed medical-decision making requires an understanding of statistics. Primarily risk-communication is necessary in the implementation of informatics or decision aids to support activities like conveying risk associated with various medical interventions like cancer treatment or screening for breast cancers, counseling patients during difficult decisions under uncertainty, or under shared medical decision making (Ancker, Senathirajah, Kukafka, & Starren, 2006; Zikmund-Fisher et al., 2008). There is some support for using graphical risk information to increase risk-avoidance behavior in people (Stone, Yates, & Parker, 1997), and reduction in biases

associated with incremental risk perceptions via pictographs for communicating side-effects of using tamoxifen for breast cancer (Zikmund-Fisher et al., 2008). However, individual differences between high and low numerates are prevalent in research pertaining to graphical vs. numerical conveyance of risk (Gaissmaier et al., 2012).

One instance from medical decision making is the need for a proper understanding of long-term risks and benefits associated with a treatment plan, which is required for patients to decide about the treatment as many treatments and diseases carry cumulative risk. A study (Fuller, Dudley, & Blacktop, 2004) illustrated that patients significantly misinterpreted, overestimated, and underestimated the risk associated with stroke over subsequent years after they received treatment for the first occurrence of stroke. The underestimation of cumulative risk has also been observed in the context of the perceived risk of sexually transmitted diseases (Knäuper, Kornik, Atkinson, Guberman, & Aydin, 2005).

Cumulative risk has also been a major concern in the case of rare but catastrophic events. De La Maza, Davis, Gonzalez, and Azevedo (2019) studied the understanding of the cumulative risk associated with flooding risk for people in the US because flood is identified as one of the major disasters in the US, but still the demand for flood insurance even with discounted premiums was relatively low compared to what it should be due to high flood risk. They suspected that this might be because of the tendency of people to ignore risk pertaining to low-probability rare events and underestimating the cumulative nature of this low probability risk as potential factors in the flood risk situation. They found the majority of individuals following additive heuristic and consequently underestimating the risk.

The scenarios mentioned before demonstrate that conveying risk is a major concern in all the uncertain environments like catastrophic events (floods, earthquakes, and tsunamis), medical diagnosis and treatment intervention, chain-smoking, the effectiveness of contraceptive methods, and so on. More importantly, calibrating people's judgment and choices to the actual cumulative risk seem to be a challenge in all the uncertain environments. However, in all the situations mentioned before, people, if taught, can calculate the cumulative risk associated with these events, because, for example, the risk of flood in year-1 is more or less independent of the risk of flood in say year-2 or -5. The same is not true in the case of storms represented via NEXRAD to pilots because one red cell is not independent of other red cells around it and the dynamics of those red cells are governed by the laws of fluid dynamics. Pilots are expected to estimate this mission-centric risk to keep their flight path safe and the risk constantly changes making it almost impossible for any pilot to calculate that risk on their own and requires complex meteorological ensemble modeling.

The goal of the current study is to convey this cumulative risk through a simulated probabilistic FRREED. In general, the cumulative risk is conveyed either in conjunctive or disjunctive probabilities (Doyle, 1997). Doyle found significant evidence that people understand disjunctive probabilities (probability of an event occurring at least once) more clearly compared to conjunctive probabilities (probability of an event never occurring). For example, pilots flying an aircraft in hazardous weather would like to know the chance of crossing the hazardous weather at least once in their flight (disjunctive). This thesis attempts to convey cumulative risk in a disjunctive manner to make pilots aware of the risk in such a hazardous situation. The next section provides an overview of the targeted research questions and the experimental design plan for this study.

## 1.8 Study Overview

The focus of the current research study is to simulate and implement a probability-based FRREED for weather displays with a primary focus on convective weather displays (especially NEXRAD) and to understand participant behavior after inclusion of FRREED in the system. The participants were asked to decide about the safety of multiple flight routes based on the available weather information and/or the FRREED. FRREED is a probability-based component that provides the probability of a particular route coming within a 20 nmi-radius-circle of hazardous weather. FRREED provides a probability, estimating the cumulative risk of the associated hazardous weather intersecting with the given flight route (at least once). The study investigates whether the participants can comprehend the uncertainty information presented to them and then make efficient use of the risk information to make better flight route safety decisions.

The experiment design comprises of four participant groups supported by varying levels of accuracy of FRREED. The “no” support condition has no FRREED (the NEXRAD only condition). The “baseline” support condition employs a FRREED whose accuracy is consistent with current capability in meteorological modeling. The “moderate” support condition employs a FRREED whose accuracy is likely at the top of what is achievable in meteorology in the near future. The “high” support display provides a level of support that is likely unachievable in an aviation weather decision-making context without significant technological innovation. The accuracy of the FRREED was manipulated using the Brier score, which is a probability score widely established for assessing the accuracy of a probabilistic forecast by using both the predicted forecast probability and the outcome of the event (Brier, 1950; Murphy & Winkler, 1977; Yates, 1990). The required baseline or moderate or high accuracy (i.e., the mean Brier score) was selected for FRREED beforehand, and the outcome of the events was available because I used historical

storm data. FRREED was simulated to produce probabilities for the forecasts from the available data. The participants in each experimental group received training in which feedback regarding their decision was presented after each trial. This was followed by a test phase in which feedback regarding the participant's decision was presented after completing (flying) the entire storm (4 trials). The division of the experiment into the training and the test phase facilitates my evaluation of how much learning and transfer took place. The experimental study is a Qualtrics plus Pavlovia (Psychopy 3 online) (Peirce et al., 2019) web-based or lab-based computer study.

The proposed hypotheses for the experimental study are the following:

**Effect of inclusion of FRREED on participants decision-making and calibration to the tool:**

- Hypothesis 1: The participants receiving probabilistic FRREED will make more accurate flight-path decisions (high proportion correct and high sensitivity) and confidence judgments (low Brier score) in order of the level of accuracy of the FRREED. The high accuracy FRREED will lead to better-calibrated participant decisions to the system, compared to the participants receiving lower support.

**Effect of learning on participants decision-making and calibration to the tool:**

- Hypothesis 2: The accuracy of the participants' flight-path decisions and confidence judgments will improve as a function of training. The transfer of learning from training to test phase is expected.
- Hypothesis 3: Performance will improve across training blocks as participants receive feedback and gain experience throughout the training phase.

- Hypothesis 4: Performance will improve across storms as participants receive feedback and gain experience throughout the test phase.
- Hypothesis 5: The unfolding behavior of a storm will improve the participants' performance on later trials when storm unfolds compared to initial ones (within-storm). I predict that in subsequent trials, as the storm moves, the participants can use that additional movement information while making flight-path decisions.

**Effect of task type on performance:**

- Hypothesis 6: Performance on double route trials (relative judgment task) will be higher than performance on single route trials (the absolute judgment task) for all four groups. This prediction is based on the findings from the area of sensation and perception, which shows that people find it easier to make judgment on relative tasks compared to absolute tasks (E. U. Weber & Johnson, 2009).

**Role of trust in automation:**

- Hypothesis 7: The participants' reliance or trust on the FRREED will change depending on the manipulated accuracy of the FRREED. I expect that reliance or trust will be positively correlated with the level of support.

## 2. METHODS

The experiment outlined in this section addresses the hypotheses presented before. The study is designed using Qualtrics and PsychoPy (Peirce et al., 2019) software for participants to make a weather-related flight-path safety decision. The experiment design was a 2 (trial type) x 2 (experiment phase) x 4 (FRREED) mixed design. Both the trial type and the experiment phase manipulations are within-subject factors, and the FRREED manipulation is a between-subjects factor. The FRREED manipulation has four levels: 1) no support (NEXRAD only), 2) low support (NEXRAD + baseline-accuracy FRREED), 3) moderate support (NEXRAD + moderate-accuracy FRREED), and (4) high support (NEXRAD + high-accuracy FRREED). The trial type manipulation has two levels: 1) a single route trial (absolute judgment) and 2) a double route trial (relative judgment). The experiment phase manipulation has two levels: 1) the training phase and 2) the test phase. The study has the following dependent variables: performance (proportion of correct weather-related decisions), trust score (trust in automation scale by Jian, Bisantz, and Drury (2000)), calibration (Brier and Brier skill score), sensitivity and bias (signal detection theory).

### 2.1 Participants

21 (Male, N= 20; Female, N= 1) pilots participated in the online version of the study advertised to them through flyers and emails sent to pilot academy listservs and posts on Reddit, pilot forums, and Facebook. Although almost 315 pilots opened the online experiment link, the attrition rate was very high, with 70% not proceeding beyond the consent form page, and others leaving in-between from the instructions to the experiment stage.

I ran 343 undergraduate students in this study in Decision Processes Lab using the online experiment management system at Georgia Institute of Technology. Participants received course credits for their participation. Only 324 participants' data (at least 70 each condition) was included

in the analysis. The discarded participants had issues with either the data file saved, or they pressed escape button in-between the study, or they skipped a lot of trials. All the participants with available response data for more than 85% of the trials were considered for the analysis due to analysis methods being sensitive to a large amount of missing data.

## **2.2 Materials**

Participants were presented with a simulated flight path in an adverse weather condition, and they had to decide whether the route was safe to continue flying based on the FAA's 20-nmi rule. NEXRAD was used as a primary meteorological product to represent hazardous weather, and it was integrated with a probabilistic FRREED for three of the four groups. Fig. 3 represents an example of a single-route trial and the decision question that the participant had to answer. The 20 nmi-radius circle (range ring) on the flight route represents the safe area around the current flight location, which was intended to serve as a guide for participants to keep clear of any red weather cells. The solid black line on the route is the projected 45 min future flight path for which participants had to make safety-decisions. The black ticks represent 15 min time intervals for the given route. The dashed line represents the aircraft's previous flight path, which is there for representation purposes. Participants were asked to assume constant altitude (cruising altitude) and constant speed (140 knots) for the entire experiment (See APPENDIX A).

The experiment consisted of a training phase followed by a test phase. Both the phases required the participants to decide the safety of flying on a given path(s) for the future 45 min by examining the given weather conditions. Fig. 3 represents a demonstration of a single route trial for the no support condition. Participants were given a pre-study questionnaire (see Appendix B and C) and a post-study feedback questionnaire (see Appendix D) at the end with the trust in

automation scale in it. All the NEXRAD trial images were extracted from past hazardous storms in the US (see APPENDIX E).

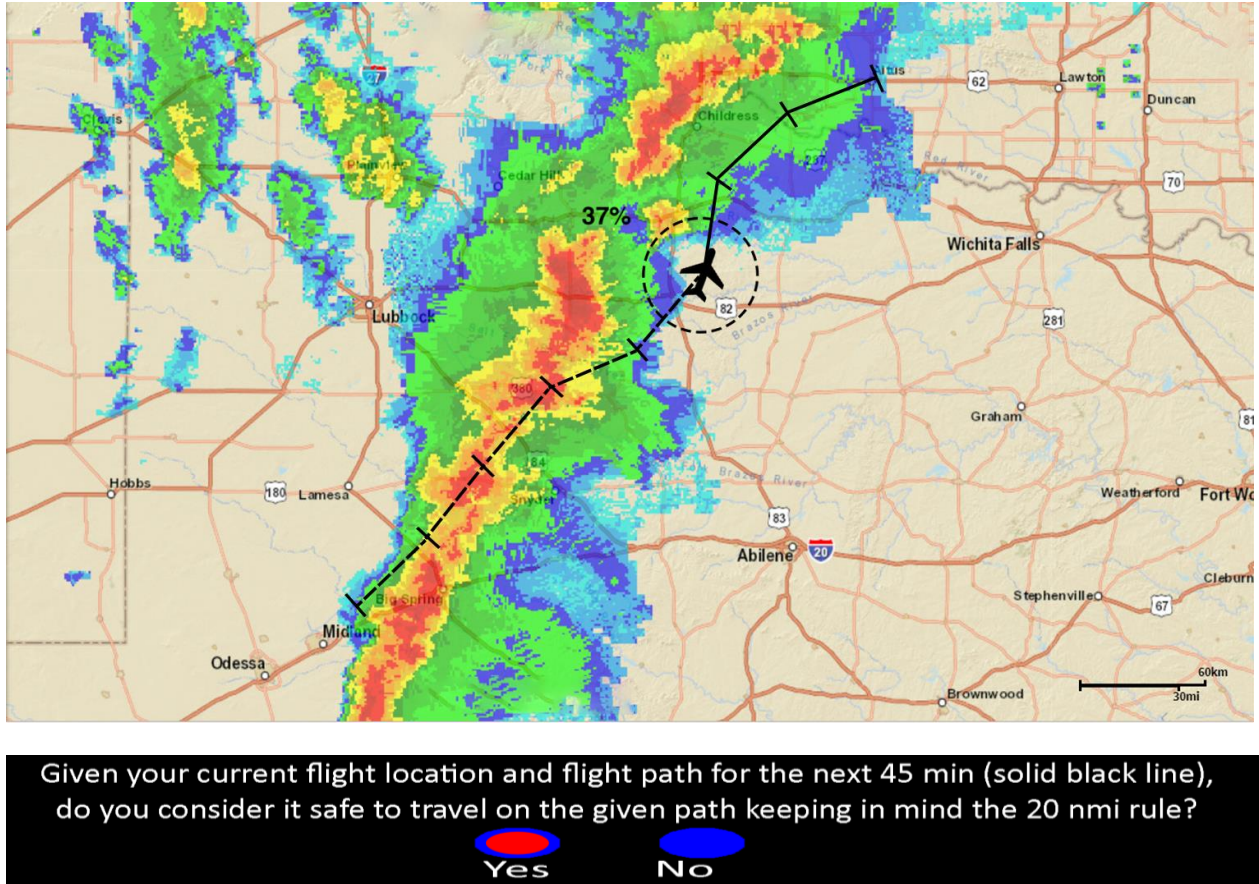


Figure 3. Single route trial with moderate-accuracy FRREED

### 2.2.1 Training and Test Phase

The training phase had 20 trials with both single and double route trials in equal proportion, presented at random. All the trials of the training phase were extracted from a different hazardous scenario (or storm). Fig. 4, represents one complete double-route trial. Before every decision trial, a loop of three NEXRAD images for every 15 min interval before the decision time were displayed once to provide the 45-minute historical weather data for the decision point. Following every decision trial, participants were asked to make confidence judgments (discussed later). Immediate

feedback was provided by three NEXRAD images (looped once) for every 15 min interval after the decision time point. This feedback was expected to facilitate the understanding of participants of the provided decision support and to enhance the transfer of learning to the test phase.

The test phase presented ten different storms with four trials each. Five of the storms were constructed as single-route trials, while the remaining five were double-route trials. Storms were presented in random order to each participant. A complete three-hour flight path was constructed through each storm. The test phase was constructed to evaluate whether the learning of participants on the probabilistic FRREED and/or NEXRAD from the training phase transferred to the more ecological test (unfolding storms) phase. Fig. 5, illustrates what participants were presented for one storm with single-route trials during the test phase. The loop of three NEXRAD images before the decision trials was displayed once to provide the historical weather data before the decision time point (initial loop shown in Fig. 4). Following this, participants were asked to make a safety decision and confidence judgment for four subsequent 45 min flight segments (trials). After the end of the fourth flight segment, participants were shown a combined feedback loop (Fig. 4) for all the four trials comprising of three images per trial for 15 min interval times.

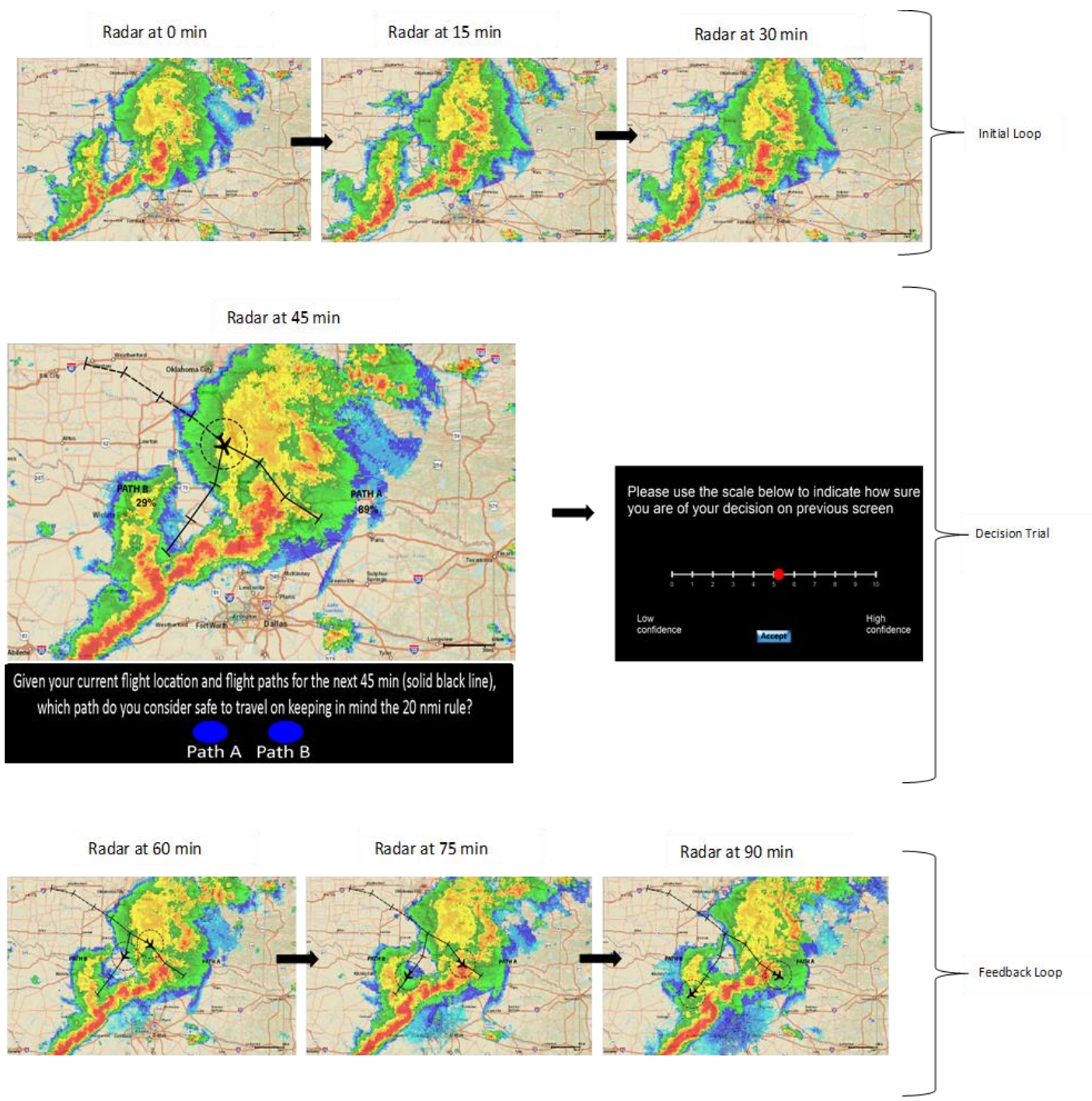


Figure 4. Elicitation for one double route trial in the training phase (with high accuracy FRREED)

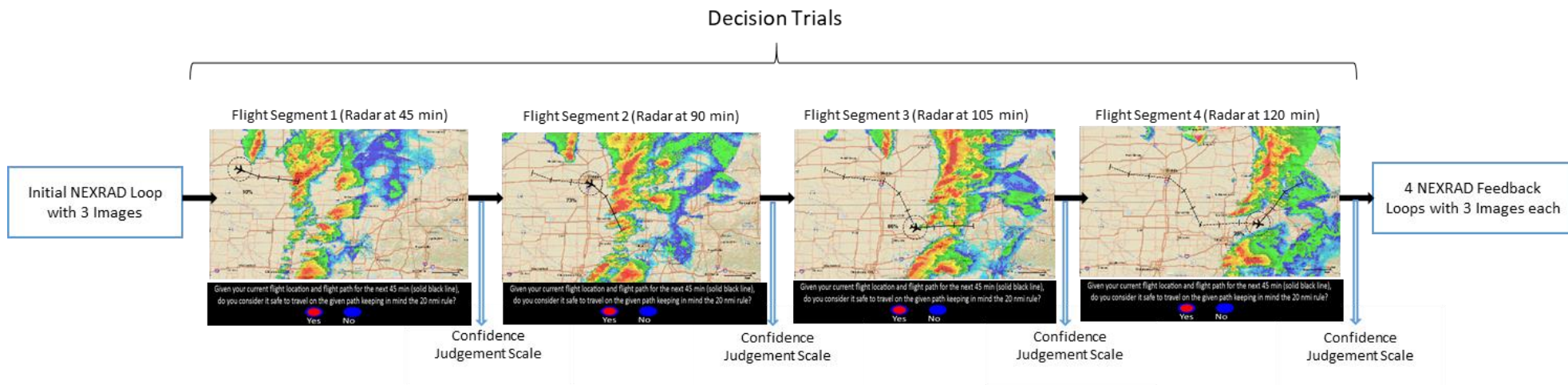


Figure 5. Elicitation for one storm with single-route trials in the test phase (with baseline accuracy FRREED)

### 2.2.2 Probabilistic FRREED

The probabilistic FRREED was simulated to have either high, moderate, or baseline accuracy as measured via Brier scores. The FRREED provides the probability that a given 45-min-route will come within 20 nmi of hazardous weather at least once along the projected flight path. The probabilities are simulated such that the Brier Score of the high accuracy FRREED is 0.04, moderate accuracy FRREED is 0.09, and baseline accuracy FRREED is 0.17. The goal was to evaluate whether the high and the moderate support condition significantly improves upon the baseline or the no support condition in terms of both decision-making and calibration to uncertainty for operators. It is feasible to simulate the probabilities from the accuracy because radar images from the past hazardous storms have been used for this experiment, and the future outcome of those storms are already known (via the available radar record of the entire storm event). Fig. 4, represents a decision trial with a high accuracy FRREED providing the probability of path A as 69% and path B as 29%. This means path A has a 69% chance of intersecting the red cells in the 20 nmi range ring during the next 45 min route shown, and path B has a 29% chance of intersecting the red cells. Ideally, if the FRREED was deterministic, the participant should always choose to go with the low probability routes. However, it is impossible to design a completely accurate system for hazardous storms due to uncertainty. So, it is important for people to account for this uncertainty in FRREED while using this for enhancing situation awareness pertaining to risk.

## **2.3 Procedure**

All the four experimental groups were tested on the same training and test phase with both single and double route trials. The manipulation defining each group was the accuracy level of the FRREED. The participants were randomly assigned to the four conditions: no support (Pilots, N=2; Students, N=73), baseline support (Pilots, N=7; Students, N=81), moderate support (Pilots, N=7; Students, N= 88), and high support (Pilots, N=5; Students, N=82). All participants went through the same radar images presented in the random order (see APPENDIX E).

I did not provide any explicit information regarding the accuracy of the NEXRAD display to the participants because such information can distort use and expectations (Barg-Walkow & Rogers, 2016), and I wanted to capture effects of experiential learning as cleanly as possible. Thus, I expected the participants to learn the accuracy of the display from the feedback provided to them. The learning of accuracy in the four conditions is evaluated through the participants' bias and sensitivity towards using the FRREED across trials. The transfer of learning is evaluated through relationship between training performance and test performance.

## **2.4 Theory for Dependent Variables**

The following sub-sections discuss the theory behind the following dependent measures: trust score, calibration (Brier and Brier skill score), and sensitivity and bias (signal detection theory) used in this study. It also discusses why the evaluation of these measures are necessary for the current study environment.

### *2.4.1 Trust in automation*

Wickens, Gordon, Liu, and Lee (1998) defined automation as reliable when it does what the human operator expects it to do. In this study, by manipulating the accuracy of the display, its reliability is also being manipulated. The trust of users' needs to be well-calibrated to the actual reliability of the system while dealing with automated systems. That means the trust in automation should always be in direct proportion to the reliability of the automation, hence, when the reliability decreases, the user should be able to rely on themselves in deciding, and should look for other sources in the environment that can provide reliable information to guide the decision. This kind of appropriate calibration prevents issues of over-trust or distrust in automation (Parasuraman & Riley, 1997), where distrust is "when the person fails to trust the automation as much as it is appropriate (Wickens et al., 1998, p. 424)" and over trust or complacency is "when the person trust the automation more than is warranted and can have serious negative consequences if automation is less than fully reliable" (Parasuraman, Molloy, & Singh, 1993; Parasuraman & Riley, 1997).

In this study, the accuracy of the display is being manipulated, and it is necessary to investigate participants' trust calibration to this manipulated accuracy of the display. Therefore, at the end of the experiment, participants were asked to respond to a 12-item Trust in Automation scale, which is developed by Jian et al. (2000). It is one of the most used scale in the Trust literature. The total score on the scale indicates the level of trust with the high scores implying overtrust on the system, and the lower scores implying distrust.

Based on the drive you just completed, please fill out the following questions where 1 = not at all and 7 = extremely.

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)
The system is deceptive. (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The system behaves in an underhanded manner. (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am suspicious of the system's intent, action, or outputs. (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am wary of the system. (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The system's actions will have a harmful or injurious outcome. (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am confident in the system. (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The system provides security. (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The system has integrity. (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The system is dependable. (9)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The system is reliable. (10)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can trust the system. (11)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am familiar with the system. (12)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 6. Trust in automation scale by Jian et al. (2000)

#### 2.4.2 Confidence Judgments and Calibration

Confidence Judgment is a process of asking participants to make a judgment on a percentage or probability score about how confident they are that their decision is correct, immediately after they respond to the task question (Efklides & Misailidi, 2010). It is optimal to take confidence judgment on a continuous scale having 5-10 categories McKelvie (1978). Less than 5 categories lead to the loss of discriminative power and validity. Yet, there is no psychometric advantage over 10 categories (or a large number of

categories). Based on this literature, a continuous scale with 10 categories is included for participants to rate their flight-path safety decision (Shown in Fig. 4). N. Weber and Brewer (2003, p. 490) state that “Calibration refers to the extent to which participants’ judgments of confidence in their decision (made on percentile scale) correspond with the actual probability that their decision was correct”. For example, if the person is 70% confident in his decision, and has a 70% probability of being accurate, then the confidence judgment can be said to be accurately calibrated to the truth. The relationship between confidence judgment and accuracy is analyzed using the probability scoring method described in the next sub-section.

### 2.4.3 Brier Score and Brier Skill Score

The most common metric for the analysis of confidence judgment is a probability score known as the Brier Score (Brier, 1950; Murphy & Winkler, 1977; Yates, 1990). The Brier score (BS) is a proper scoring rule that provides a measure of the accuracy of confidence judgments:

$$BS = \frac{1}{N} \sum_{T=1}^N (c_t - o_t)^2$$

Equation 1: Brier Score

The Brier score is described by Eq. 1, where  $N$  is the total number of probability or confidence assessments,  $c_t$  is the  $t^{\text{th}}$  confidence judgment, and  $o_t$  is the outcome index for the  $t^{\text{th}}$  confidence judgment. If the event occurs, then  $o_t = 1$ , and if the event does not occur, then  $o_t = 0$ . Thus, the Brier score is the average squared deviation between the confidence of the decision-maker and the outcome index (Brier, 1950; Murphy & Winkler, 1977;

Yates, 1990). The lower the Brier score for a set of predictions, the better the predictions are calibrated (i.e., less error in predictions).

Another measure, the Brier Skill Score (BSS) acts as an overall measure of the system's performance (Wilks, 1995). It measures the relative skill of a forecast above a reference (baseline) forecast using Eq. 2.

$$BSS = \frac{(1 - BS)}{BS_{ref}}$$

Equation 2: Brier Skill Score

It ranges from minus infinity to 1. BSS= 0 implies no skill compared to the reference forecast. BSS=1, imply perfect score

Hence, BS is a measure of the accuracy of predictions or forecasts. BSS is a measure of the proportion of improvement in accuracy over the reference (Dance, Ebert, & Scurrah, 2010).

#### 2.4.4 *Signal Detection Theory*

Most decisions in real-life take place in the presence of some uncertainty because of inherent ambiguity in the system and/or time pressure associated with decision-making. Signal detection theory (SDT) provides a method to model decision making under uncertain situations (Swets, Dawes, & Monahan, 2000). It has its roots in sensory experiments, where there is a need to detect the presence of a signal in a noisy environment.

For example, in the task of identifying a tumor from a patient's CT scan, there is always some uncertainty associated with the doctor's response because of the complexity

of the task (Heeger, 2003). SDT accounts for four possible outcomes possible in an uncertain situation (shown in Table 2): Hit (tumor present and doctor responds “yes”), Miss (tumor present and doctor responds “no”), False Alarm (tumor absent and doctor responds “yes”), and Correct Rejection (tumor absent and doctor responds “no”). According to SDT, the response to a stimulus depends both on the individual’s sensitivity to the stimulus in the presence of noise and on the individual’s decision criterion (or bias).

Table 2. Outcomes table for signal detection theory

	<b>Response: Yes</b>	<b>Response: No</b>
<b>Stimuli present</b>	Hit (H)	Miss (M)
<b>Stimulus absent</b>	False alarm (FA)	Correct rejection (CR)

Sensitivity or Discriminability Index ( $d'$ ) is the ability to tell that two signals are different.

$$d' = Z_{FA} - Z_{Hit}$$

Equation 3: Sensitivity or Discriminability Index

Response Bias or Criterion (C) is an individual’s implicit decision threshold above which they respond “yes” to the presence of a signal and below which they respond “no” to the presence of a signal.

$$C = -(Z_{FA} + Z_{Hit})/2$$

Equation 4: Response Bias or Criterion

### 3. RESULTS

This section will address the hypotheses presented in the introduction section through both the analysis description and findings. The sub-sections are presented in terms of the dependent variables measured. Bonferroni corrections to alpha levels are used when required to control for familywise Type 1 error.

#### 3.1 Performance (Proportion Correct) Trends across Conditions

Participant performance is defined as a binary outcome with correct response coded as 1 and incorrect response coded as 0 for every trial for each participant. Therefore, the mean performance closer to 1 is better and closer to 0 is poor. Fig. 7 shows the trend in the increase in mean performance for student data with an increase in the level of support. The trend indicates that student participants did rely on the FRREED to improve their performance over the no-support condition (NEXRAD only). In the same Fig 7, the trend for mean performance for pilots' data is not as good as the students except for the high support condition. However,  $N=23$  for pilots' data and  $N=324$  for student participants' data. Hence, I am going to make conclusions going forward cautionary, but there doesn't seem to be any large discrepancy between expert-novice performance that I can detect. Also, the students performed nearly as close to the pilots in two conditions, outperformed in one, and are slightly below in the no support condition. Hence, with the acknowledged limitation, it is safe to make conclusions about the usability of FRREED based on student data from this point onwards. It is also interesting to mention here that  $N=2$  for no-support condition in pilots, and we observed the highest attrition rate for pilots in this condition. Therefore, we only had data from outperformers (or experts) in this group, which provides

further evidence on why our student participants lagged in this condition. From this point onwards, results will only focus on student participants due to lack of power for pilot population.

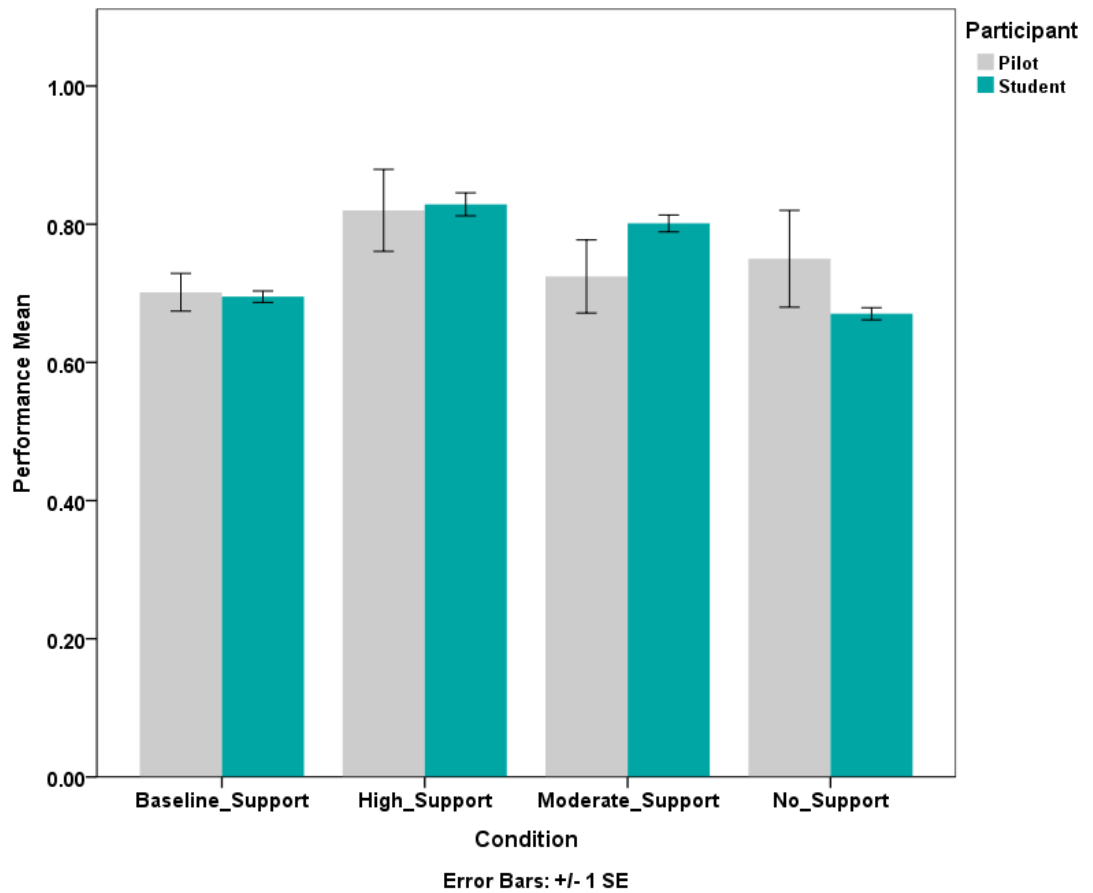


Figure 7. Effect of level of support on performance

### 3.2 Sensitivity and Bias Analysis

A decision support in this context should facilitate people's ability to discriminate if their path will be intersected by hazardous weather (sensitivity). Although it is important to separate accuracy or sensitivity from the effects of a decision support on people's risk

aversion (response thresholds), both aspects of decision making are important to evaluation (Swets et al., 2000).

The class of Generalized Estimating Equations (GEE) for categorical data can be configured to estimate sensitivity and bias parameters within a single model specification (via the use of a probit link function). Using the probit link function, parameter estimate for each factor manipulation represent the changes in decision threshold (in beta units), and parameter estimates for each factor manipulation crossed with “truth,” the actual occurrence of the storm, represents changes in discriminability (in d-prime units). In Table 3, examining the variables that interacted with truth shows that discriminability improved as the level of support increased (accuracy of tool) (mean trend shown in Fig. 8). Discriminability also improved more ( $p < 0.0001$ ) in the test phase compared to the training phase with the increase in the level of support (Truth  $\times$  Experiment Phase  $\times$  Condition), and it also improved more in double route trials ( $p < 0.01$ ) compared to single route trials with the increase in the level of support (Truth  $\times$  Trial Type  $\times$  Condition). These results are in support of hypotheses 1, 2, and 6.

Although there weren't any *a priori* predictions concerning decision thresholds, it is still important to evaluate. The decision thresholds significantly increased as the level of decision support increased (mean trend shown in Fig. 9). This is consistent with the idea that the decision thresholds are positively correlated with the accuracy of the FRREED; decision makers adopt more conservative thresholds as the accuracy of a system increases.

Table 3. Summary statistics for Type 3 generalized estimating equations analysis, displaying flight-path safety decisions regressed on information characteristics and task outcome

<b>Independent Variable</b>	<b>Df</b>	<b>Wald Chi-Square Statistic</b>
Condition	3	9.30*
Trial Type	1	13.51***
Truth	1	945.16****
Experiment Phase	1	45.74****
Condition × Trial Type	3	7.91
Condition × Truth	3	89.14****
Condition × Experiment Phase	3	5.68
Trial Type × Truth	1	106.70****
Trial Type × Experiment Phase	1	104.18****
Truth × Experiment Phase	1	89.49****
Condition × Trial Type × Truth	3	15.32**
Condition × Trial Type × Experiment Phase	3	15.49**
Condition × Truth × Experiment Phase	3	32.46****
Trial Type × Truth × Experiment Phase	1	3.24
Condition × Trial Type × Truth × Experiment Phase	3	4.29

\*p<0.05

\*\*p<0.01

\*\*\*p<0.001

\*\*\*\*p<0.0001

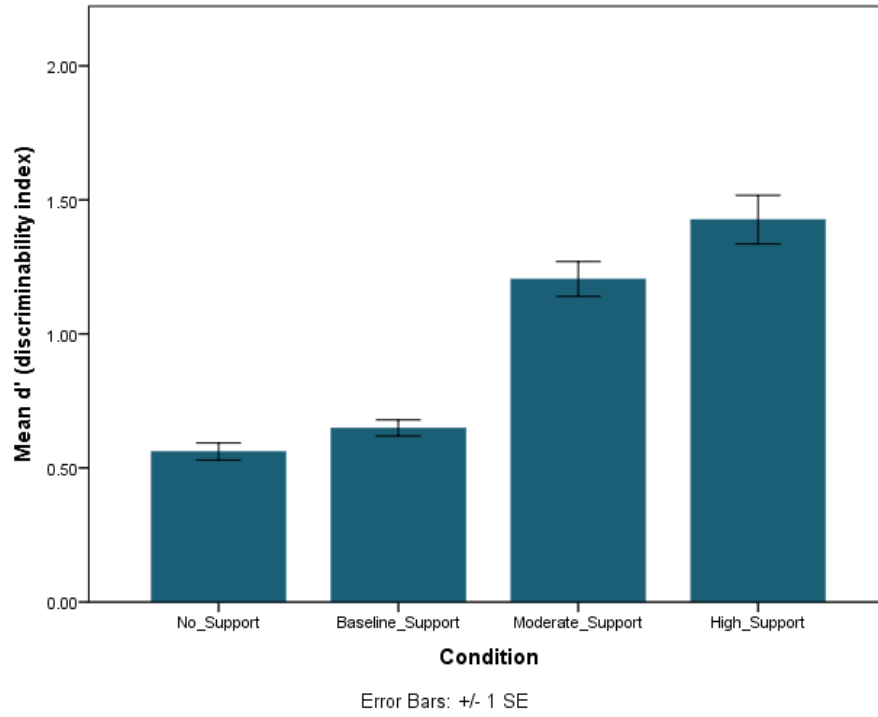


Figure 8. Trend in mean discriminability across conditions

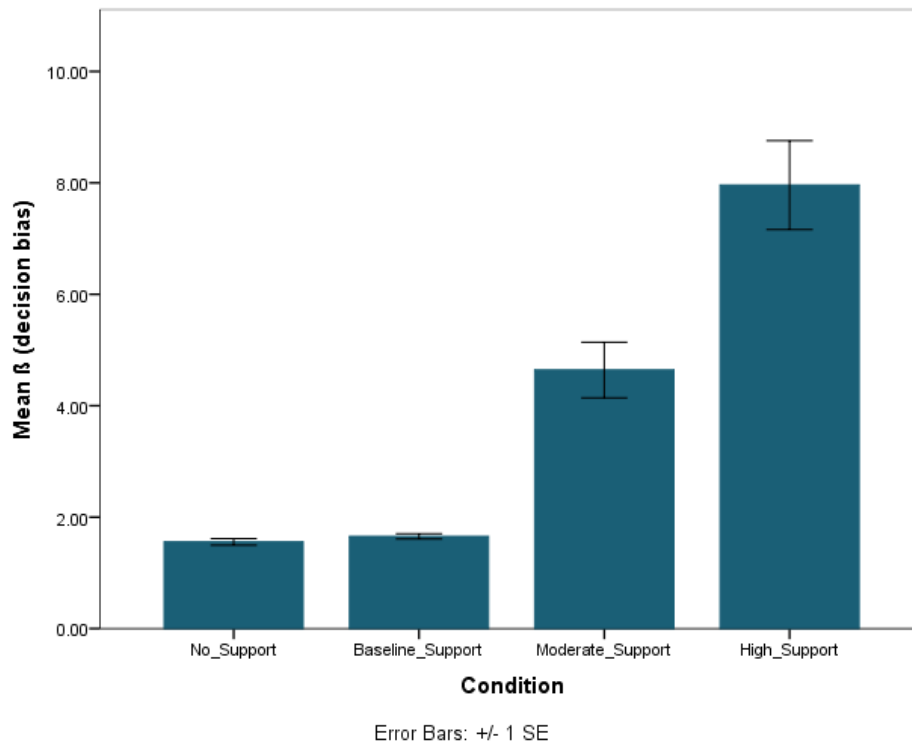


Figure 9. Trend in mean bias across conditions

### 3.3 Confidence Judgment Calibration

Brier Score (BS) measurements are used to test the calibration of participant confidence judgments. There is a significant decrease in the mean Brier Score for participants as the accuracy of the decision support increases (Table 4). Although the calibration of the participants increased as the accuracy of the decision support increased, the mean BS is still lower than the corresponding BS for the FRREED itself. Thus, the participants are not perfectly calibrated to the accuracy of the FRREED and do not utilize all the information provided by the FRREED to its potential. Another important observation to note is mean BS for NEXRAD-only condition is approximately equal to mean BS of 0.25 for chance performance.

I calculated the Brier Skill Score (BSS) to scale the improvement in accuracy provided by the FRREED compared to the control (no support -NEXRAD only) condition. The higher the BSS means, the greater the proportion of improvement in calibration over the control condition. Table 4, shows that BSS is highest for the high support condition, and all the FRREED conditions improved participants' forecasting skill above the control condition.

For the statistical analyses of Brier Scores, I employed a Generalized Linear Model for repeated measures to test the differences in Brier Scores between conditions, trial type, and phase (Table 4). This analysis indicate that Brier Scores significantly decreased (better calibration) as the level of support increased. Moreover, the Brier Scores were significantly lower in the relative judgment trials compared to the absolute judgment trials, and in the

test phase compared to the training phase. These findings yield support for the calibration predictions of hypotheses 1, 2, and 6.

Table 4. Group wise Mean Brier Score and Brier Skill Scores

Condition	Mean BS for FRREED	Mean BS for Participants	Mean BS for Participants		Mean BSS for Participants (ref-no support)
			Training Phase	Test Phase	
No Support	-	.24	.26	.24	-
Baseline Support	0.18	.22	.25	.22	0.08
Moderate Support	0.09	.18	.21	.16	0.26
High Support	0.04	.14	.18	.12	0.41

Table 5. Summary statistics for Type 3 generalized linear model analysis, displaying Brier Score regressed on information characteristics

Independent Variable	Df	Wald Chi-Square Statistic
Condition	3	102.01****
Trial Type	1	79.88****
Experiment Phase	1	124.39****
Condition × Trial Type	3	12.47
Condition × Experiment Phase	3	9.86*
Trial Type × Experiment Phase	1	10.11**
Condition × Trial Type × Experiment Phase	3	20.65****

\*p<0.05

\*\*p<0.01

\*\*\*p<0.001

\*\*\*\*p<0.0001

### 3.4 Learning Within-Subject

The following sub-sections addresses all the learning hypotheses including transfer of learning, within-phase learning, and unfolding-order learning.

#### 3.4.1 *Transfer of Learning*

I predicted that the performance in the training phase (overall, first-half, and second-half) will predict the performance in the test phase. To evaluate the transfer of learning predictions, I calculated the Kendall's tau-b correlation between performance in the training and the test phases for each condition. There was a positive correlation between the overall training phase performance and the test phase performance (Table 6). The correlation was significant for all the FRREED conditions and the strength of correlation increased as the accuracy of the FRREED increased. The correlation was not significant for the no support condition, indicating that NEXRAD-only was insufficient to facilitate transfer. Table 6 also reports performance correlations between first-half (10 trials) or second-half (10 trials) of the training phase with the test phase. There is similar but weaker trend for these correlations. Correlations were also calculated for Brier Score between the training and the test phases (Table 6) to check how the calibration in the training predicts calibration in the test. The BS correlation was significant for all the conditions and the strength of correlation increased as the accuracy of the support increased. Hence, the proposed probabilistic support facilitated transfer to a more ecological environment. These results yield support for the transfer of learning predictions of hypothesis 2.

Table 6. Kendall's Tau and Pearson's Correlations [ $\tau_b$  ( $r$ )] matrix between training and test phase

		Overall Test Phase Performance				
		No Support	Baseline Support	Moderate Support	High Support	Total
<b>Training Phase Performance</b>	<b>Overall</b>	.14 (.16)	.20** (.29)	.47**** (.61)	.66**** (.77)	.45**** (.61)
	<b>First-Half</b>	.08 (.08)	.23* (.28)	.28**** (.37)	.52**** (.61)	.33**** (.43)
	<b>Second Half</b>	.10 (.14)	.06(.11)	.44**** (.56)	.67**** (.77)	.42**** (.55)
		Overall Test Phase Brier Score				
<b>Training Phase Brier Score</b>	<b>Overall</b>	.25** (.38)	.26**** (.38)	.53**** (.69)	.60**** (.76)	.48**** (.71)
	<b>First-Half</b>	.14 (.23)	.23** (.299)	.34**** (.44)	.45**** (.57)	.34**** (.49)
	<b>Second Half</b>	.25** (.36)	.16* (.29)	.53**** (.69)	.62**** (.77)	.48**** (.69)

\*p<0.05

\*\*p<0.01

\*\*\*p<0.001

\*\*\*\*p<0.0001

### 3.4.2 Training Phase Learning

In hypothesis 3, I predicted that performance should improve across training blocks due to learning through feedback. To test hypothesis 3, 20 trials of the training phase were divided into 5 training blocks with 4 trials each (in order of trial presentation). I employed a Generalized Linear Model for repeated measures to test the differences in performance between conditions, training blocks, and scenario (Table 7). I included scenario in the

model for all the within-subject learning hypotheses (3,4, and 5), because all the hazardous storms were different in strength, movement, and uncertainty. The analysis indicates that the main effect of training blocks is significant, i.e., performance significantly increased across training blocks (Fig. 10). Therefore, providing support for the learning prediction of hypothesis 3. There is no significant interaction between condition and training blocks, indicating that the performance over subsequent training blocks doesn't change significantly with the change in the level of support.

Moreover, as expected, the main effect of scenario is significant, indicating significant differences in performance for all 20 trials of the training phase (Figure 11). Also, there is significant interaction between training blocks and scenario, indicating that learning is not consistent across all training blocks. This is likely due to scenario difficulty (e.g., scenarios - 4, 25). The interaction between condition and scenario is also significant, indicating that easier scenarios were easier and difficult were less difficult in order of the support level received by participants.

Table 7. Summary statistics for Type 3 generalized linear model analysis, displaying performance in the training phase regressed on information characteristics

<b>Independent Variable</b>	<b>Df</b>	<b>Wald Chi-Square Statistic</b>
Condition	3	25.31****
Scenario	19	856.89****
Training Blocks	4	24.36****
Condition × Training Blocks	12	13.24
Condition × Scenario	57	96.02**
Scenario × Training Blocks	76	100.79*

\*p<0.05

\*\*p<0.01

\*\*\*p<0.001

\*\*\*\*p<0.0001

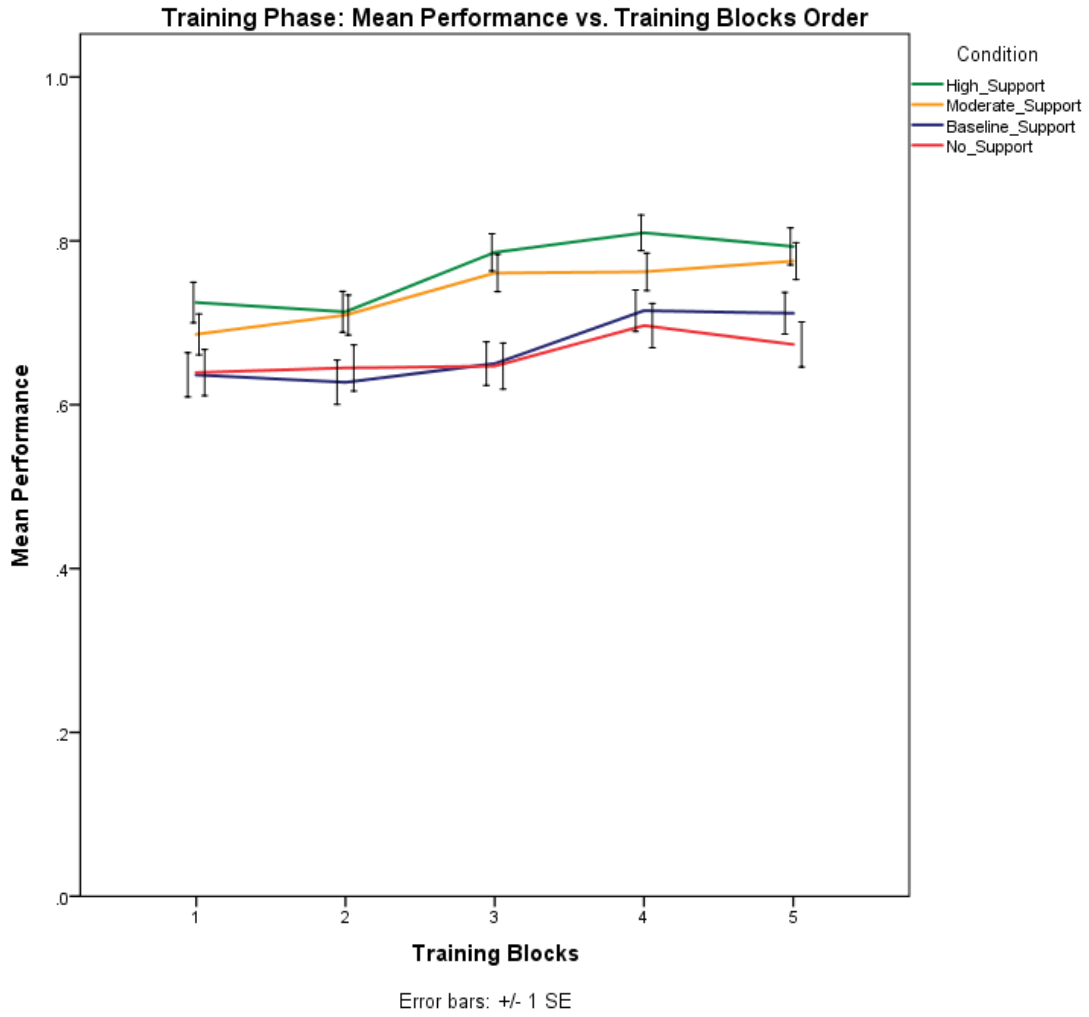


Figure 10. Trend in mean performance for subsequent training blocks across conditions

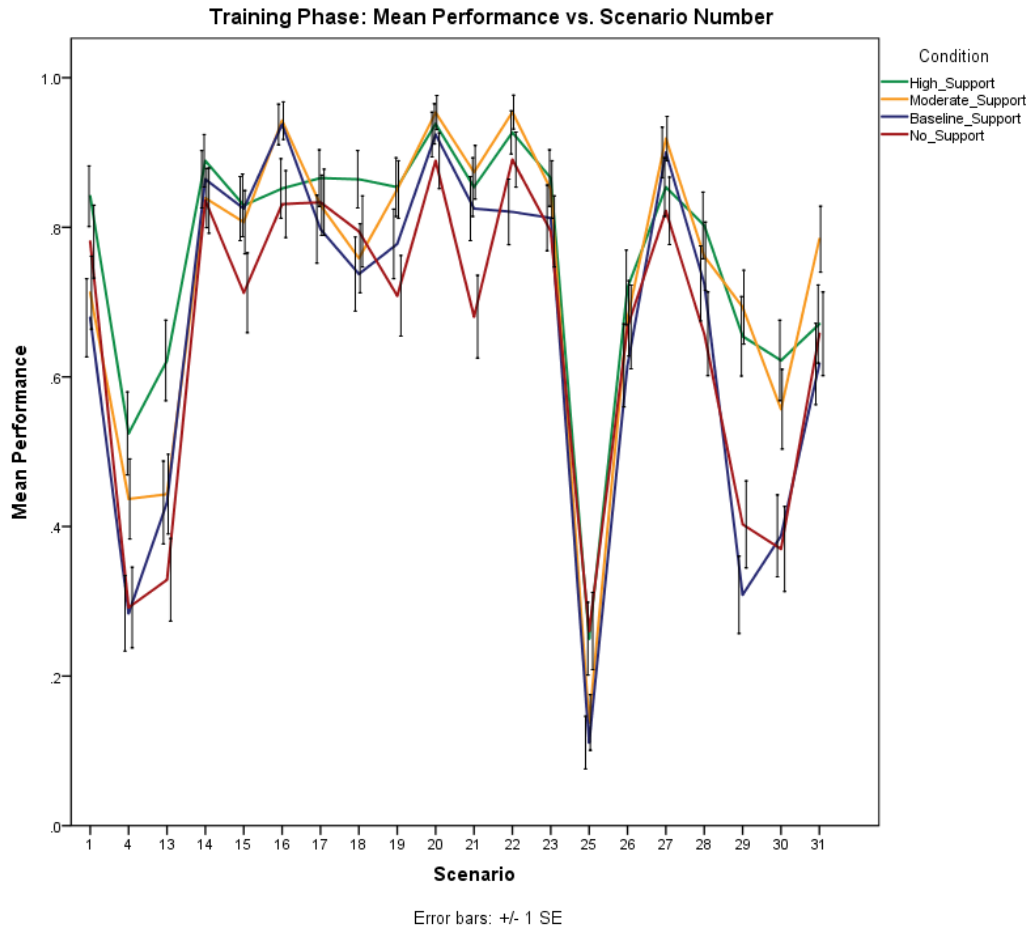


Figure 11. Trend in mean performance for different scenarios across conditions

### 3.4.3 Test Phase Learning

In the test phase, each scenario had 4 different unfolding trials in which the same storm was evolving in time. To address my predictions regarding learning within the test phase, I employed a Generalized Linear Model for repeated measures to test the differences in performance between conditions, storm presentation order, unfolding trial order, and scenario (Table 8). The main effect of storm presentation order is significant, providing

support for learning predictions with subsequent storms (Fig. 12). Moreover, the interaction between condition and storm presentation order is also significant, indicating that performance improved within the test phase, and it improved more for the high support condition compared to the other support levels. Therefore, providing support for the learning predictions for the test phase in hypothesis 4. Also, the scenarios differed in difficulty level evident from the significant main effect of scenario (Fig. 13).

For the unfolding trial order, I predicted that performance will improve when going from trial 1 to trial 4 for all scenarios in the test phase due to learning and understanding a particular storm. The main effect of unfolding trial order and the interaction between unfolding trial order and condition is significant (Table 8). However, the trend from trial 1 to trial 4 is completely opposite of what was expected. The performance decreases from 1 to 4, and it decreases less for high support compared to other support levels (Fig. 14). Hence, there was no support for the unfolding predictions in hypothesis 5. However, consistent decreasing performance trend across all support levels can be attributed to the additional NEXRAD loop available before the initial trial but was not available later as the storm unfolded.

Table 8. Summary statistics for Type 3 generalized linear model analysis, displaying performance in the test phase regressed on information characteristics

<b>Independent Variable</b>	<b>Df</b>	<b>Wald Chi-Square Statistic</b>
Condition	3	91.88****
Scenario	9	535.85****
Storm Presentation Order	9	18.24*
Unfolding Trial Order (within each storm)	3	183.78****

Table 9. *Continued*

Independent Variable	Df	Wald Chi-Square Statistic
Condition × Scenario	27	82.25****
Condition × Storm Presentation Order	27	51.60**
Condition × Unfolding Trial Order	9	37.21****
Storm Presentation Order × Unfolding Trial Order	27	39.35
Scenario × Unfolding Trial Order	27	638.77****

\*p<0.05

\*\*p<0.01

\*\*\*p<0.001

\*\*\*\*p<0.0001

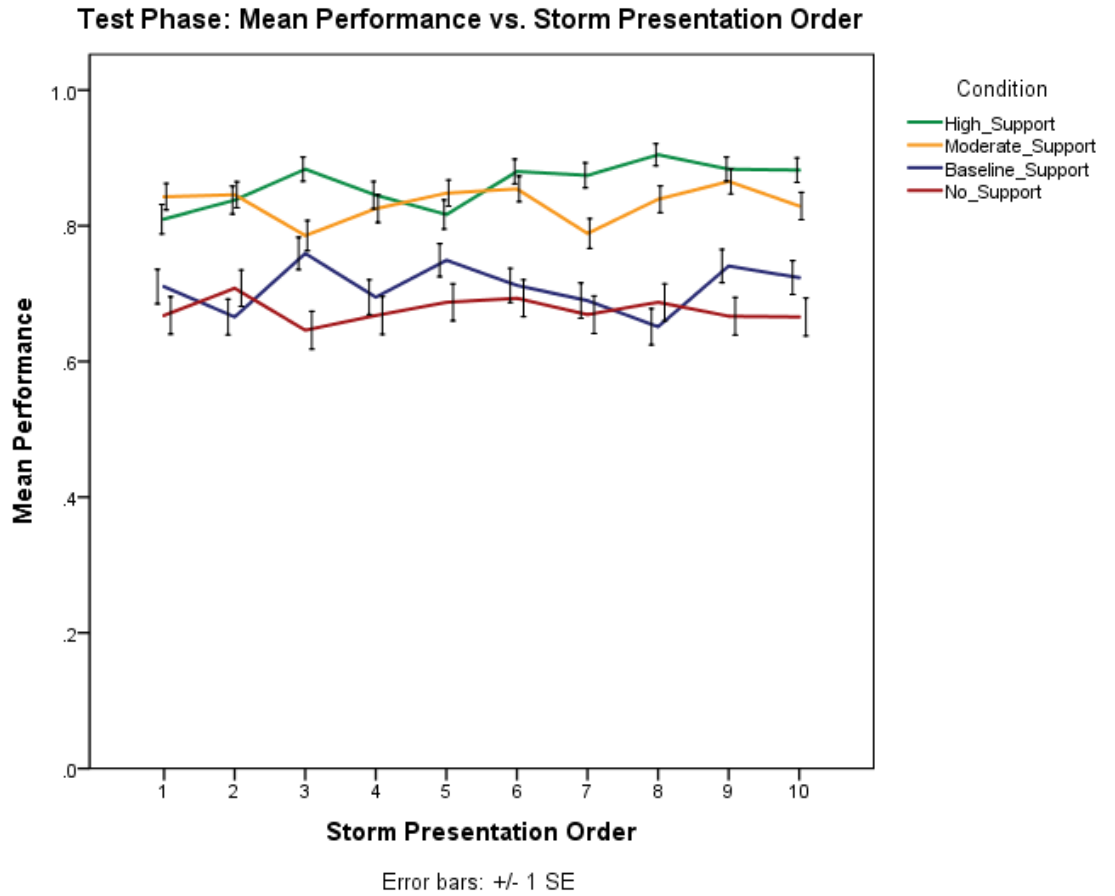


Figure 12. Trend in mean performance for subsequent storms presented across conditions

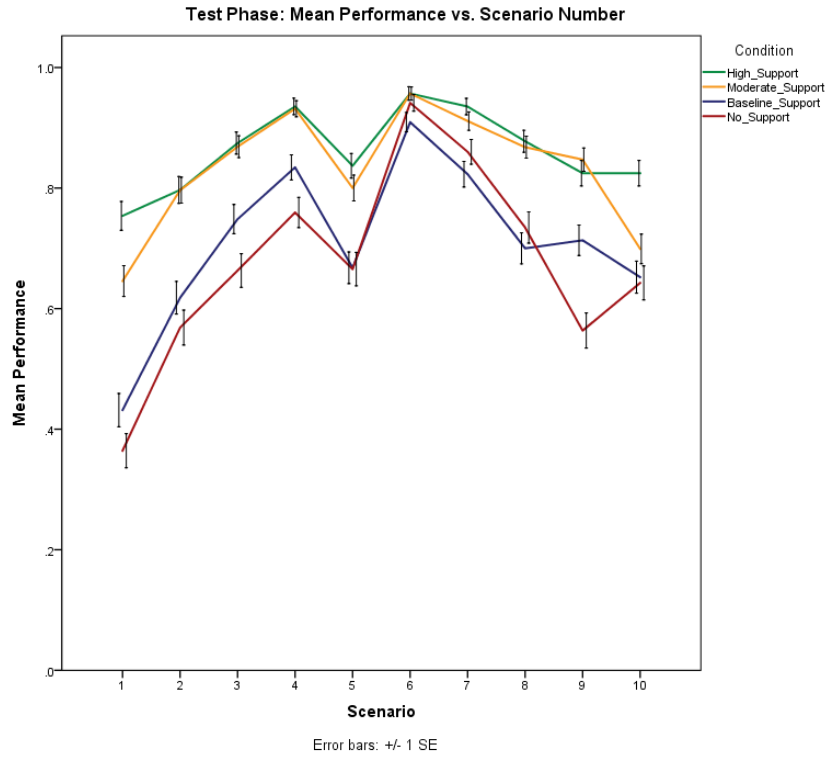


Figure 13. Trend in mean performance for different scenarios across conditions

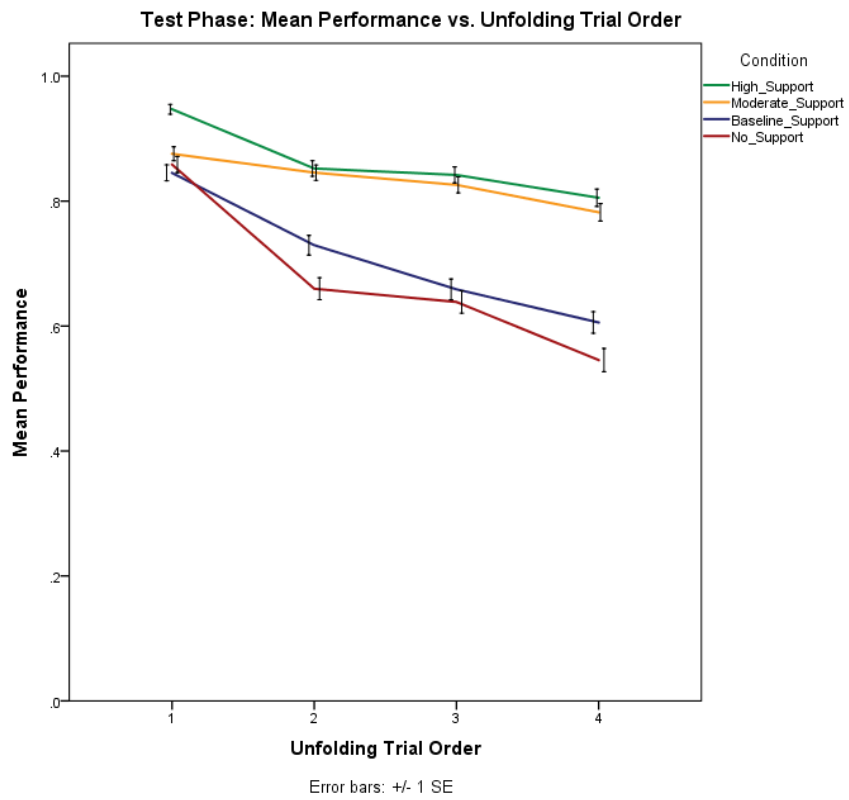


Figure 14. Trend in mean performance for unfolding trials across conditions

### 3.5 Trust in Automation

This section addresses hypothesis 7 concerning trust (or reliance) in the system. Prior to testing this hypothesis, an evaluation of how the component trust (positively worded items) and the component distrust (negatively worded items) scores are related is done for the entire data. The component trust and the component distrust scores are found to be significantly negatively correlated,  $r = -0.494$  and  $p < 0.0001$ . The trust score is calculated by reverse scoring the first five items in the trust scale (distrust questions) and then taking an average of all 12 responses. Fig. 15 shows the group-wise mean trust scores.

A one-way ANOVA indicates a significant mean difference in trust score across support level,  $F = 6.524$ , and  $p = 0.0001 < 0.05$ . This is followed by Tukey's HSD test for post-hoc comparisons to test where the mean difference exists. However, the post-hoc comparisons only indicate significant mean differences in trust scores between the baseline support and the moderate support conditions, and the baseline and the high support conditions,  $p = 0.003 < 0.008$ . Hence, there is only partial evidence found for hypothesis 7. This is counter-intuitive because of the significant calibration of participants' confidence judgment to the accuracy of the support conditions.

found in Section 3.3. However, several of the items on the trust in automation scale do not seem particularly relevant to the context of a probabilistic decision support tool. This makes sense in that the trust scale items are designed for higher levels of automation or warning systems, which provides deterministic information to the users. Therefore, I decided to do some exploratory analysis by picking up the three most relevant items to the probabilistic decision-support context and testing the trust score on them separately. The

item about confidence in the system, trust in the system, and dependability of the system were selected and analyzed for any mean differences across groups through one-way ANOVA followed by post-hoc comparisons. A one-way ANOVA indicates significant mean difference in trust score across support level for the confidence item ( $F= 3.756$   $p=0.011<0.05$ ), the dependability item ( $F= 5.194$   $p=0.002<0.05$ ), and the trust item ( $F=6.233$   $p=0.0001<0.05$ ). However, none of the post-hoc comparisons were significant for the confidence item. The mean score was significantly different between the baseline and the moderate support conditions for the dependability item ( $MD= -0.82$ ,  $p=0.001<0.008$ ). The mean score was significantly different for the no support and the baseline support conditions ( $MD= -0.71$ ,  $p=0.005<0.008$ ), and the baseline support and the moderate support conditions ( $MD= -0.739$ ,  $p=0.002<0.008$ ) for the trust item. Therefore, these exploratory results were similar to what I found for the overall trust score previously.

These findings led to further exploratory analysis of how the trust scale and its individual items correlate with the Confidence Judgment and Brier Scores measured previously. This comes from the idea that both the previously measured Confidence Judgment (CJ) and Brier Scores (BS) are logical substitutes for the measure of trust (reliance) in the system. Kendall's tau and Pearson correlation were calculated between overall trust score, the component trust score, the component distrust score, and the individual trust scale items with Brier Scores and Confidence Judgments (Table 9). The component distrust score and the score from five distrust items are expected to be negatively correlated with the CJ and positively correlated with the BS. The overall trust score, the component trust score, and the score from seven trust items are expected to be positively correlated with the CJ and negatively correlated with the BS. The overall trust

score and the component trust score obtained from 7 trust items are found to be significantly correlated with CJ ( $p < 0.0001$ ) and BS ( $p < 0.0001$ ) in the expected direction for only the high support condition. A similar trend is also observed in all seven trust items. Six out of the seven trust items have significant correlation with BS and CJ for the high support condition only, and most of the correlations for the other support conditions are not significant. The component distrust score (from five items) and five individual distrust items are not correlated with CJ and BS except for a small significant correlation between the BS and the component distrust score (for high support), the BS and the wary item (for no support), and the BS and the deceptive system item (for moderate support). The correlation matrix from Table 9 further supports my claim that the trust scale does not fit the context of a probabilistic decision support tool, especially with all the distrust items. Also, the trust scale seems to be correlated with the CJ and BS for only the high support condition indicating the sensitivity of the scale to the accuracy of the support provided.

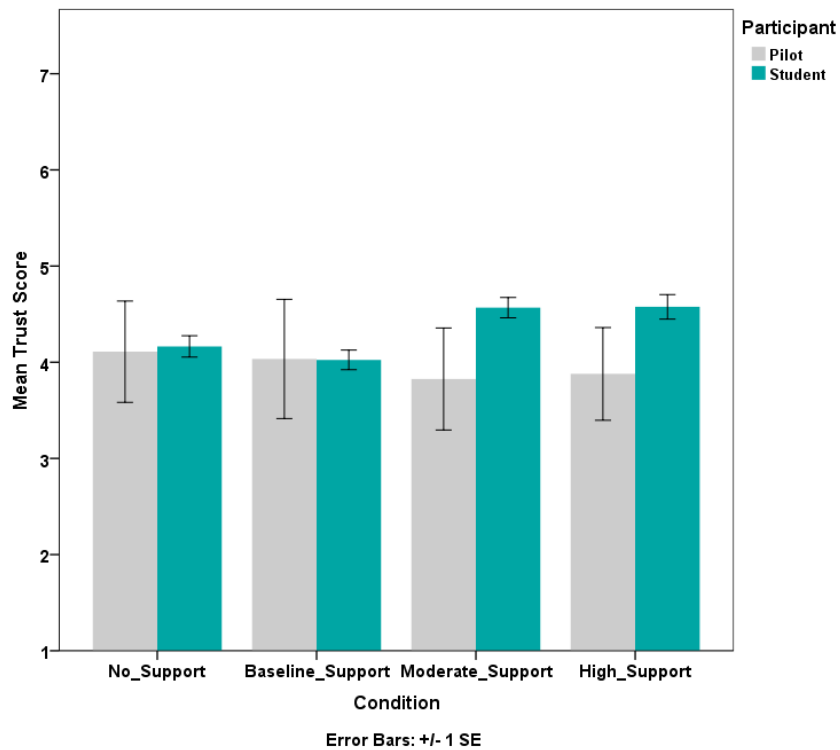


Figure 15. Mean trust in automation score for each condition

Table 10. Kendall's Tau and Pearson's correlations [ $\tau_b$  ( $r$ )] matrix for trust scale items, and Confidence Judgments and Brier Scores

$\tau_b$ ( $r$ )	No Support		Baseline Support		Moderate Support		High Support		Total	
	CJ	BS	CJ	BS	CJ	BS	CJ	BS	CJ	BS
<b>Overall Trust Score</b>	.12 (.15)	-.04 (.03)	.02 (.02)	-.11 (-.16)	.10 (.08)	-.16 (-.26*)	.29**** (.34)	-.36**** (-.53)	.16 (.20)	- .23**** (-.36)
<b>Component Distrust Score</b>	-.01 (-.05)	.05 (.04)	.001 (-.004)	.1 (0.17)	-.04 (-.05)	.14 (.25*)	-.17 (-.19)	.27** (.39)	-.06 (-.08)	.17**** (.23)
<b>Component Trust Score</b>	.16 (.21)	.05 (.09)	.048 (.08)	-.12 (-.17)	.05 (.03)	-.11 (-.19)	.30**** (.38)	-.40**** (-.55)	.17**** (.25)	- .22**** (-.33)
<b>Deceptive System</b>	-.05 (-.13)	.09 (.10)	.05 (.06)	.12 (.18)	-.15 (-.19)	.23* (.31)	-.02 (-.02)	.14 (.24)	-.06 (-.08)	.18 (.24)
<b>Underhanded</b>	-.04 (-.05)	.08 (.07)	.02 (.05)	.06 (.13)	.20 (.33*)	-.04 (-.07)	.04 (.05)	.09 (.18)	.03 (.08)	.07 (.08)
<b>Suspicious</b>	-.07 (-.15)	.08 (.15)	-.07 (-.05)	.15 (.18)	.14 (.21)	.19 (.23)	.17 (.22)	.08 (.07)	.01 (.03)	.13* (.16)
<b>Wary</b>	-.24* (-.27)	.04 (.01)	-.11 (-.13)	-.05 (-.10)	.09 (.22)	.06 (.05)	-.13 (-.21)	.20 (.24)	-.10 (-.11)	.08 (.13)
<b>Harmful</b>	.16 (.18)	.06 (.04)	-.02 (.03)	.06 (.11)	-.14 (-.07)	.24 (.21)	-.20 (-.35)	.36** (.45)	-.06 (-.05)	.22**** (.27)
<b>Confident</b>	.23** (.31)	-.001 (.04)	.00 (.005)	-.11 (-.13)	.13 (.08)	-.19* (-.23)	.39**** (.46)	-.46**** (-.61)	.20**** (.23)	- .23**** (-.33)
<b>Security</b>	.17 (.18)	.06 (.15)	.02 (-.004)	-.10 (-.12)	.16 (.14)	-.17 (-.23)	.31**** (.40)	-.44**** (-.58)	.19**** (.22)	- .21**** (-.31)
<b>Integrity</b>	.08 (.10)	.22* (.31)	.04 (-.003)	-.12 (-.15)	.07 (.09)	-.20 (-.16)	.24** (.30)	-.36**** (-.47)	.14** (.14)	-.15** (-.21)

Table 9. *Continued*

$\tau_b(r)$	No Support		Baseline Support		Moderate Support		High Support		Total	
	CJ	BS	CJ	BS	CJ	BS	CJ	BS	CJ	BS
<b>Dependable</b>	.13 (.17)	.11 (.16)	-.04 (-.10)	-.002 (.04)	.20* (.24)	-.15 (-.24)	.25* (.31)	-.34*** (-.45)	.16** (.18)	-.15** (-.23)
<b>Reliable</b>	.14 (.21)	.04 (.06)	.1 (-.03)	-.02 (.01)	.0 (.09)	-.19* (-.28)	.21* (.26)	-.31**** (-.43)	.13** (.15)	- .18**** (-.25)
<b>Trust</b>	.15 (.14)	.01 (.08)	.04 (.06)	-.10 (-.12)	.16 (.16)	-.19 (-.20)	.26** (.31)	-.34**** (-.47)	.18**** (.20)	- .22**** (-.30)
<b>Familiarity</b>	.02 (.08)	-.21* (-.32)	.17* (.24)	-.23** (-.29)	-.04 (-.03)	-.08 (-.11)	.13 (.09)	-.17 (-.20)	.10* (.13)	- .21**** (-.26)

## 4. DISCUSSION

This study explored a predictive FRREED providing the cumulative risk associated with a weather hazard and asking participants to make flight-path safety judgments. The research investigated the transfer of learning, trust calibration to the accuracy of the system, judgment calibration, and the appropriate representation of uncertainty required to improve performance and decision-making. The results show that operators did rely on the FRREED to improve their performance over the no-support condition. The Brier Skill Score for the high support condition in reference to the no support condition was 0.41. Therefore, showing that the level of performance of the operators improved in terms of both calibration and resolution as support increased in its accuracy. These results demonstrate that the operator performance is not acceptable in absolute terms with the use support level with accuracy commensurate with the current capability of meteorological probability forecasts (no support, BSS=0, or baseline support, BSS=0.07). Therefore, suggesting a need for similar research with pilots followed by the validation of training regimes for the safe introduction of high-accuracy probabilistic decision-aids in general aviation cockpits.

A myriad of performance and calibration measures show significant differences between different levels of support conditions very consistently. However, the operators are still not completely calibrated to the accuracy of the tool (Table 4), and there is further room for improvement here. The strong transfer of learning effects from the training phase to the test phase (unfolding storms) in FRREED conditions demonstrated success in understanding and using probabilistic risk estimates. The learning effect implies that

operators may not be able to use probabilistic risk estimates to their full potential without training. The learning effects within the training and the test phase shows the importance of enhancing learning via multiple trials- and scenario-based training, which is suggestive that experience-based training modules could be effective method for training operators on probabilistic weather risk estimates.

I predicted that performance was going to increase in unfolding trials as more information about the storm movement would become available to the operator as the storm unfolded. However, I found an opposite trend that shows the decrements in performance when you move farther away in the storm. Perhaps this might be because operators find it challenging to extract cues about storm movement in the absence of the NEXRAD initial loop after the first unfolding trial. The performance decrement trend might be due to the lack of feedback after every unfolding trial, which might have facilitated the storm movement's understanding. This finding emphasizes a need for constant feedback and constant past data to make real-time decisions about hazardous weather.

The results from the trust in automation scale indicate that operators were neutral toward the support tool. Hence, the trust scores for operators don't seem to calibrate with their actual performance in the task. This is non-intuitive because all the performance and calibration measures showed that people did rely on using the FRREED. Interestingly, the exploratory analysis shows that the level of trust in the FRREED was only calibrated to the operator's objective performance when the FRREED was highly accurate. Thus, the non-intuitive trend in trust scores can be attributed to the structure of the trust scale itself, which seems suitable for a higher level of automation instead of an aid that renders uncertainty via numerical probability. Our results also indicate the possibility of people calibrating

their trust in the system only when the system is highly accurate. Also, the design of the experiment might be affecting the trust calibration (Stuck & Walker, 2020), because of the inability of student participants to perceive the risk associated with intersecting a storm while flying.

#### **4.1 Implications**

Overall, the results of this thesis project have implications for new interventions in the aviation industry. It has implications for display designers in terms of introducing products with probabilistic weather information into the market place as well as introducing the regulatory regime for the entry of such products into the cockpit. These results guide the development of probabilistic decision aids that people can comprehend easily. Thus, it addresses some of the issues faced by earlier attempts by Matthews and DeLaura (2010), Kronfeld (2003), and others (Table 1), in introducing qualitative probabilistic risk information in cockpits which pilots continued to underestimate. The results of the transfer of learning and within-phase learning can help in developing and improvising training programs for weather displays. This finding should reinforce the need for general aviation pilot training to include dedicated modules for weather displays and the interpretation of probabilistic weather products, which is only a negligible portion of the training they receive currently.

The results also contribute to basic research in judgment and decision making about how to facilitate people's understanding of cumulative risk and uncertainty. It specifically addresses the long-believed notion in decision-making literature that lay people find probability format difficult to understand and account for in the decisions compared to the

more commonly known frequency format. There is a multitude of psychological evidence showing how people have difficulty in understanding probabilities compared to frequencies (Slovic, 1997; Tversky & Kahneman, 1974). Our results clearly demonstrate the participants' ability to understand risk estimates in probability format and calibrate themselves to the accuracy of those risk estimates. Hence, further providing evidence that people can use probability formats effectively (c.f., the work of Joslyn and Nichols (2009) showing the understanding of probabilistic rain forecasts amongst non-experts).

The results have implications for some of the previous challenges demonstrated in conveying cumulative risk in various domains like breast-cancer medication risk (Zikmund-Fisher et al., 2008) , contraception methods related risk (De La Maza et al., 2019), stroke recurrence risk (Fuller et al., 2004), flood risk (De La Maza et al., 2019), etc. mentioned before. The results show that by scenario- and feedback-based training it is possible to calibrate people's judgment and decision-making to the numerical risk estimates and the inherent uncertainty in the complex mission-centric task of flight-route judgments. Therefore, this can be extended to other less complex domains in which the events (like red cells on NEXRAD) are not completely correlated with each other.

## **4.2 Limitations**

This study had two major goals. First, to find the best way to convey cumulative risk in an uncertain environment through probabilistic risk estimates and second, to determine what level of accuracy of these probabilistic displays will facilitate learning and performance. To support these two goals, the experiment consisted of four groups supported by varying accuracy of the risk estimates, and every group of participants had to

go through two phases to facilitate the transfer of learning. Although the intent was to make the test phase more like a continuous flying task, only static NEXRAD images were provided to the participants. Also, participants had very little control to go over the initial loop multiple times. An actual general aviation cockpit NEXRAD display allows users to go over the past loop of NEXRAD images again and again. Yet, it was restricted in the current experiment to once to control the NEXRAD loop from confounding the effect of the risk estimates on operator performance. It was also done to restrict the length of the study by keeping every task timed.

Another major limitation is that the participant pool comprised of a convenience sample (undergraduate students) who have little to no experience with flying or dealing with hazardous weather in any capacity. Although all the efforts were made to run the experiment online on a web browser, recruiting professional pilots to participate in the study was the biggest hurdle. However, if students can understand the quantitative estimate of risk and can calibrate themselves to the accuracy of the estimates, I predict that when pilots will be provided with the decision aid of this high accuracy, they will be able to use it efficiently.

### **4.3 Future Research**

Future research should build upon the findings of this study by performing ensemble modeling on big weather data from the past, which can help in getting actual real-time cumulative risk values. Along with this, the study should be conducted with a high-fidelity flight simulator and professional pilots to evaluate how the current findings extend to the environment in which it needs to be implemented. Also, the next steps should

be to evaluate the ways to build operator trust in the kind of probabilistic systems evaluated in this study. The trust in automation results also highlights the need to build a trust in automation scale that is catered to predictive decision-support systems. Future research should also investigate ways to reach the accuracy of the high support condition implemented in this study, which requires technological innovations not present at this time.

## **APPENDIX A. PARTICIPANT INSTRUCTIONS**

The upcoming pages show the exact representation of the instructions that were displayed to the participants except for example radar images.

### **Basic Instructions**

You are about to view a series of radar images depicting storms. Your task is to determine if your flight path will come within 20 nautical miles (nmi) of hazardous weather over the course of 45 min of flying. You will also be asked to judge how confident you are in your flight-path decision. More details for this part of the experiment will be provided later.

Following your response(s), you will receive visual feedback about whether your flight path would have come within 20 nmi of hazardous weather.

### **Flight Conditions**

When evaluating the flight paths assume your aircraft is traveling at 140 knots and that the speed of your aircraft remains constant. It means your aircraft will move a distance of 35 knots in every 15 min. Pictorial depiction of the flight path is shown later.

For your altitude, assume that you are flying at a constant altitude throughout the experiment. Basically, assume that you cannot change altitude to escape the hazardous weather.

## Radar Image Explanation

The radar images that you will see are NEXRAD (Next-Generation Radar) composite maps. An example is shown below.

The composite maps represent data from multiple National Weather Service (NWS) radars and show patterns of precipitation. The precipitation in the NEXRAD map is color-coded, as shown in the scale below. The scale units, dBZ, represent the different echo intensities (reflectivity) measured by the radar during each elevation scan. "Reflectivity," designated by the letter Z, is the amount of transmitted power returned to the radar receiver and covers a wide range of signals, from very weak to very strong. Because of this, a decibel (or logarithmic) scale dBZ is used. The scale below represents dBZ values when the radar is in precipitation mode.

The scale of the dBZ values is also related to the intensity of rainfall. Typically, light rain is occurring when the dBZ value reaches 20. The higher the dBZ, the stronger the rain rate.

Hazardous weather is represented by the red activity (50+ dBZ) in our images, and it will be your job to determine if your flight path will come within 20 nmi of any red over the course of 45 min.

The radar images represent actual historical storms.



Your flight path is indicated by the black line. The solid black line represents the 45 min segment for which you are going to make a decision. The dashed line represents the aircraft's previous flight path, which is there for presentation purposes. The circle (range ring) represents all the area within a 20 nautical mile radius of your current position. When you are assessing whether or not your flight path(s) will come within 20 nmi of any red activity, you can use this circle as a guide. If the circle intercepts any red as it travels along the flight path line, your aircraft will have come within 20 nmi of hazardous weather. The tick marks along the line at every 15 min are there to help you project how far you would be traveling at your current speed of 140 knots. Each tick mark represents 35 nmi, which is the distance you will travel every fifteen minutes. Remember that the radar shown represents the current radar reflectivity.

- Make a prediction for the next 45 min of flight (black solid flight path)
- Speed: 140 knots

### **Flight-Path Risk Value**

The flight-path risk value located on the flight-path in the image is the probability that the flight-path will come within 20 nmi of red activity at least once. For example, a flight-path risk value of 17% means that 17 out of 100 times you flew this flight path you would travel within 20 nmi of red activity at least once. This value was derived for the purposes of this study and is not currently available on any operational forecast product. Note that the storms selected for this study are hazardous.

## Summary

To summarize, given an image like the one below, based on the current NEXRAD weather you will need to determine if the given flight path will keep your aircraft at least 20 nmi away from any red activity. You will be asked to answer the following two questions: (1) *Given your current flight path(s) and flight location, do you consider it safe to travel on the given path keeping in mind the 20 nmi rule?* and (2) *How confident are you in your decision?*

Your aircraft's flight info is:

- **Speed: 140 knots**
- **Altitude: Constant Cruise Altitude**

## Feedback

You will be provided with visual feedback on whether or not your aircraft intersected red activity after you have made your decision(s) at the conclusion of each trial. The visual feedback will illustrate the aircraft traveling along the future 45-minute flight path from one 15-min tick mark to the next, allowing you to discern whether or not your flight-path evaluation was accurate.

That concludes the instructions. When you are ready to begin viewing the radar images, please request the experimenter to run the program for you. You will have to evaluate 30-weather situations.

**Please come back to this page for the post-experiment questions and feedback after you finish the experiment.**

## APPENDIX B. PRE-STUDY QUESTIONNAIRE FOR STUDENT PARTICIPANTS

Do you have any prior flying experience? Please provide details.

Have you had any training in weather interpretation (for example, courses in meteorology)?

Yes

No

# APPENDIX C. DEMOGRAPHICS QUESTIONNAIRE FOR PILOTS

## Demographics

Please indicate your gender.

Male

Female

Other

Please indicate your race/ethnicity.

White

Hispanic or Latino

Black or African American

Asian

Native Hawaiian or Other Pacific Islander

American Indian or Alaska Native

Other

## Experience

What certifications do you currently hold? Please select all that apply.

Student Pilot License

Sport Pilot License

Recreational Pilot License

Private Pilot License

Commercial Pilot License

Airline Transport Pilot License

Other

Do you currently hold an instrument rating?

Yes

No

Do you currently hold an instructor rating?

Yes

No

Do you currently hold a Pilot-In-Command (PIC) Type Rating?

Yes

No

What airplane class ratings do you currently hold? Please select all that apply.

Multi-engine Land

Single-engine Land

Multi-engine Sea

Single-engine Sea

None

Other (Please list below)

Please roughly estimate your total number of flight hours.

Please indicate the region within which you have operated most frequently during the preceding 6 months (indicate more than one region if necessary):

**Northeast - New England**

Maine  
New Hampshire  
 Vermont  
Massachusetts  
Rhode Island  
Connecticut

**Northeast - Middle Atlantic**

New York  
Pennsylvania  
New Jersey

**Midwest - West North Central**

North Dakota  
South Dakota  
 Nebraska  
Kansas  
Minnesota  
Iowa  
Missouri

**South Atlantic**

Delaware  
Maryland  
District of Columbia  
 Virginia  
West Virginia  
North Carolina  
South Carolina  
Georgia  
Florida

**South - West South Central**

Oklahoma  
 Texas  
Arkansas  
Louisiana

**West - Mountain**

Idaho  
Montana  
Wyoming  
 Nevada  
Utah  
Colorado  
Arizona  
New Mexico

**West- Pacific**

Alaska  
 Washington  
Oregon  
California  
Hawaii

**International**

Please estimate the number of times that you have been forced to divert in order to avoid adverse weather:

Please estimate the number of hours (total) experience using weather radar systems in the conduct of a flight:

Have you had any training in weather interpretation other than basic pilot training (for example, courses in meteorology)?

Yes

No

## Weather Experience

Please indicate how often do you use the following weather resources.

	Never	Occasionally	Often
AC (Severe Wx Outlook)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
AIRMET / SIGMET	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Charts, Convective outlook	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Charts, Prog.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Charts, Radar (NEXRAD)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Charts, Radar summary	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Charts, Surface analysis	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Center Weather Advisory (ATC)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Charts, Weather depiction	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
FA (18-hr area forecast)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
FD (winds/temps aloft forecast)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
METAR	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
PIREP	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Satellite	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
TAFs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Onboard Radar	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mobile Applications (e.g., ForeFlight, Garmin, WingX)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>


Please identify the level to which you agree with the following statements.

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I feel comfortable flying in adverse weather conditions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am confident in my ability to navigate adverse weather conditions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I try to avoid flying in adverse weather.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I rely on weather radar systems to assist my navigation in and around weather systems.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I trust the information displayed on weather radar systems.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

## APPENDIX D. FEEDBACK QUESTIONNAIRE

This section shows the questions that were displayed after the trust in automation scale

How did you use the following information to evaluate the flight paths:

	How often did you use these components?			Was this component difficult to use or interpret?			
	Always	Sometimes	Never	Difficult	Moderately Difficult	Moderately Easy	Easy
Ownship icon	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Flight path tick marks	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Radar reflectivity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
 Flight Path Risk	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What challenges did you encounter while using the provided information?

What weather information would you have liked to have had for evaluating your flight pa

## APPENDIX E. NEXRAD STIMULI

**Stimuli Folder Link:** [Experiment Stimuli](#)

**Password:** 02262020\_Masters\_Thesis\_Parmar

## REFERENCES

- Ahlstrom, U. (2015). Weather display symbology affects pilot behavior and decision-making. *International Journal of Industrial Ergonomics*, 50, 73-96. doi:10.1016/j.ergon.2015.09.008
- Ahlstrom, U., Ohneiser, O., & Caddigan, E. (2016). Portable weather applications for general aviation pilots. *Human factors*, 58(6), 864-885.
- Alam, S., Abbass, H., & Barlow, M. (2006). Multi-objective ant colony optimization for weather avoidance in a free flight environment. *The Artificial Life and Adaptive Robotics Laboratory, University of New South Wales, Canberra, Australia*.
- Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, 13(6), 608-618.
- Barg-Walkow, L. H., & Rogers, W. A. (2016). The effect of incorrect reliability information on expectations, perceptions, and use of automation. *Human factors*, 58(2), 242-260.
- Bass, E. J., & Minsk, B. (2001). *Toward integrating real-time and forecast weather hazards information for improved pilot decision making*. Paper presented at the Proceedings of the Eleventh International Symposium on Aviation Psychology.
- Beringer, D. B., & Ball, J. D. (2004). *The effects of NEXRAD graphical data resolution and direct weather viewing on pilots' judgments of weather severity and their willingness to continue a flight*. Retrieved from
- Boyd, D. D. (2017). In-Flight Decision-Making by General Aviation Pilots Operating in Areas of Extreme Thunderstorms. *Aerosp Med Hum Perform*, 88(12), 1066-1072. doi:10.3357/AMHP.4932.2017
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1), 1-3.
- Burgess, M. A., & Thomas, R. P. (2004). The effect of NEXRAD image looping and national convective weather forecast product on pilot decision making in the use of a cockpit weather information display.
- Busquets, A. M., Ruokangas, C. C., & Kelly III, W. E. (2005). Advanced Weather Awareness and Reporting Enhancements.
- Campbell, S. E., & Alexander, A. (2016). *The Effect of Cockpit-Based Convective Weather Notifications on Pilot Decision Making*. Paper presented at the 16th AIAA Aviation Technology, Integration, and Operations Conference.

- Dance, S., Ebert, E., & Scurrah, D. (2010). Thunderstorm strike probability nowcasting. *Journal of Atmospheric and Oceanic Technology*, 27(1), 79-93.
- De La Maza, C., Davis, A., Gonzalez, C., & Azevedo, I. (2019). Understanding cumulative risk perception from judgments and choices: an application to flood risks. *Risk analysis*, 39(2), 488-504.
- Doyle, J. K. (1997). Judging Cumulative Risk 1. *Journal of Applied Social Psychology*, 27(6), 500-524.
- Dutcher, J. W., & Doiron, G. M. (2008). *Weather risk management through a systematic approach to the investigation of weather events*. Paper presented at the Proceedings of the ISASI 2008 Conference, Halifax.
- Efklides, A., & Misailidi, P. (2010). *Trends and prospects in metacognition research*: Springer Science & Business Media.
- Elgin, P. D., & Thomas, R. P. (2004). An integrated decision-making model for categorizing weather products and decision aids.
- FAA. (2010). Weather-Related Aviation Accident Study 2003–2007.
- FAA. (2013). *Federal Aviation Administration: Advisory Circular*. Retrieved from [https://www.faa.gov/documentlibrary/media/advisory\\_circular/ac%2000-24c.pdf](https://www.faa.gov/documentlibrary/media/advisory_circular/ac%2000-24c.pdf)
- Fuller, R., Dudley, N., & Blacktop, J. (2004). Older people's understanding of cumulative risks when provided with annual stroke risk information. *Postgrad Med J*, 80(949), 677-678. doi:10.1136/pgmj.2004.019489
- Gaissmaier, W., Wegwarth, O., Skopec, D., Müller, A.-S., Broschinski, S., & Politi, M. C. (2012). Numbers can be worth a thousand pictures: individual differences in understanding graphical and numerical representations of health-related information. *Health Psychology*, 31(3), 286.
- Gibbons, W., Holst, T., & Jonsson, J. (2013). Understanding Convective Weather Forecast Uncertainty Needs of ATM. *General Dynamics, DOT/FAA/DTFAWA-10-00029, Task Order #6*.
- Gigerenzer, G., Hertwig, R., Van Den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). “A 30% chance of rain tomorrow”: How does the public understand probabilistic weather forecasts? *Risk Analysis: An International Journal*, 25(3), 623-629.
- Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*: Cambridge university press.
- Heeger, D. (2003). Signal Detection Theory. Retrieved from <http://www.cns.nyu.edu/~david/handouts/sdt/sdt.html>

- Hunter, D. R., Martinussen, M., & Wiggins, M. (2003). Understanding how pilots make weather-related decisions. *The International Journal of Aviation Psychology*, 13(1), 73-87.
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53-71.
- Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: applied*, 18(1), 126.
- Joslyn, S. L., Nadav-Greenberg, L., Taing, M. U., & Nichols, R. M. (2009). The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision. *Applied Cognitive Psychology*, 23(1), 55-72. doi:10.1002/acp.1449
- Joslyn, S. L., & Nichols, R. M. (2009). Probability or frequency? Expressing forecast uncertainty in public weather forecasts. *Meteorological Applications*, 16(3), 309-314. doi:10.1002/met.121
- Knäuper, B., Kornik, R., Atkinson, K., Guberman, C., & Aydin, C. (2005). Motivation influences the underestimation of cumulative risk. *Personality and Social Psychology Bulletin*, 31(11), 1511-1523.
- Knecht, W. R. (2016). Current NEXRAD Cannot Reliably Enable Safe Flight Around Heavy Weather. *The International Journal of Aviation Psychology*, 26(1-2), 46-61. doi:10.1080/10508414.2016.1234936
- Kronfeld, K. M. (2003). Enhanced Weather Radar (EWxR) System.
- LeClerc, J., & Joslyn, S. (2015). The cry wolf effect and weather-related decision making. *Risk Anal*, 35(3), 385-395. doi:10.1111/risa.12336
- Matthews, M., & DeLaura, R. (2010). *Assessment and Interpretation of En Route Weather Avoidance Fields from the Convective Weather Avoidance Model*. Paper presented at the 10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference.
- McKelvie, S. J. (1978). Graphic rating scales—How many categories? *British Journal of Psychology*, 69(2), 185-202.
- Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(1), 41-47.

- Muthard, E. K., & Wickens, C. D. (2002). *Change detection after preliminary flight decisions: Linking planning errors to biases in plan monitoring*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty Forecasts Improve Decision Making Among Nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3), 209-227. doi:10.1518/155534309x474460
- NCWF. (2007). *National Convective Weather Forecast: Product Description Document*. Retrieved from <http://www.nws.noaa.gov/infoservicechanges/NCWFv2pdd.pdf>
- NOAA WEBSITE. Retrieved from <https://gis.ncdc.noaa.gov/maps/>
- NTSB. (2014). *General aviation: Identify and communicate hazardous weather*. Retrieved from National Transportation Safety Board:
- Orasanu, J., Ames, N., Martin, L., & Davison, J. (2001). Factors in aviation accidents: decision errors. *Linking expertise and naturalistic decision making*, 209.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The international journal of aviation psychology*, 3(1), 1-23.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2), 230-253.
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., . . . Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, 51(1), 195-203.
- Slovic, P. (1997). Public perception of risk. *Journal of environmental health*, 59(9), 22-25.
- Sobieralski, J. B. (2013). The cost of general aviation accidents in the United States. *Transportation Research Part A: Policy and Practice*, 47, 19-27.
- Spirkovska, L., & Lodha, S. K. (2002). AWE: Aviation weather data visualization environment. *Computers & Graphics*, 26(1), 169-191.
- Stone, E. R., Yates, J. F., & Parker, A. M. (1997). Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of experimental psychology: applied*, 3(4), 243.
- Stuck, R., & Walker, B. (2020). The importance of incorporating risk into the study of human-automation trust. *In Press*.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological science in the public interest*, 1(1), 1-26.

- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157), 1124-1131.
- Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annu Rev Psychol*, 60, 53-85. doi:10.1146/annurev.psych.60.110707.163633
- Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology*, 88(3), 490.
- Wickens, C. D., Gordon, S. E., Liu, Y., & Lee, J. (1998). An introduction to human factors engineering.
- Wiegmann, D. A., Goh, J., & O'Hare, D. (2002). The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather. *Human factors*, 44(2), 189-197.
- Wiggins, M. (2005). The interpretation and use of weather radar displays in aviation. *Aviation Research Investigation Report*). Canberra, ACT: Australian Transport Safety Bureau.
- Wilks, D. S. (1995). Statistical Methods in the Atmospheric Sciences. *International Geophysics Series*, 59(Academic Press).
- Willows, R., Reynard, N., Meadowcroft, I., & Connell, R. (2003). *Climate adaptation: Risk, uncertainty and decision-making. UKCIP Technical Report: UK Climate Impacts Programme.*
- Wu, S.-C., Ligda, S. V., Lim, V., Comerford, D. A., & Johnson, W. W. (2010). Designing Flight Deck Displays to Mitigate Weather Impacts under Trajectory-Based Operations.
- Yates, J. F. (1990). *Judgment and decision making*: Prentice-Hall, Inc.
- Yuchnovicz, D. E., Novacek, P. F., Burgess, M. A., & Stokes, A. F. (2001). Use of a data-linked weather information display and effects on pilot navigation decision making in a piloted simulation study.
- Zikmund-Fisher, B. J., Ubel, P. A., Smith, D. M., Derry, H. A., McClure, J. B., Stark, A., . . . Fagerlin, A. (2008). Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. *Patient education and counseling*, 73(2), 209-214.