INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2022.2121036

Taylor & Francis
Taylor & Francis Group

Exploring the Impact of Visualization Design on Non-Expert Interpretation of Hurricane Forecast Path

Barbara Millet[a], Sharanya J. Majumdar[b], Alberto Cairo[a], Brian D. McNoldy[b], Scotney D. Evans[c], and Kenneth Broad[b]

[a]School of Communication, University of Miami, Coral Gables, FL, USA; [b]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL, USA; [c]School of Education, University of Miami, Coral Gables, FL, USA

ABSTRACT

Hurricane forecast graphics have the challenging task of communicating information about spatial and temporal uncertainty. Although forecasting accuracy has improved, the popular track forecast cone or "Cone of Uncertainty" graphic, produced by the National Hurricane Center, is poorly understood by the general public. A better understanding of the forecast can potentially assist in timely decisions and life-saving actions. This study evaluates the impact of visualization design, tropical cyclone characteristics, subjective numeracy, and subjective graphicacy on visual attention to and user interpretation of hurricane forecast graphics. Forty-three non-expert participants completed forecast path estimates for nine tropical cyclones, comparing their responses when using the National Hurricane Center's cone of uncertainty graphic and two alternative forecast visualizations. Results show that design modifications did not alter visual attention patterns or improve interpretations. Results also indicate that subjective numeracy, subjective graphicacy, and tropical cyclone characteristics, in combination, influence estimates of hurricane forecast tracks. The findings from this study inform redesign efforts of hurricane risk communication products.

1. Introduction

When tropical cyclones (TCs) including hurricanes threaten, people access information about them from different sources (Broad et al., 2007; Dash & Gladwin, 2007; Huang et al., 2012). They rely on television, radio, websites, mobile applications, and informal social networks to evaluate risk and make decisions. Media sources communicate the risk of hurricanes through a variety of forms, both visual and non-visual. Frequently, media show the possible future path of the center of the TC accompanied by a "cone of uncertainty" (COU), a product created by the National Hurricane Center (NHC), officially named the "Tropical Cyclone Track and Watch/Warning Graphic," and first unveiled in 2002 (see Figure 1).

The COU graphic provides the track forecast of the center of the storm, together with an estimate of track forecast uncertainty, and shows areas under a watch or warning. Technically, the cone of uncertainty represents the area within which the center of the TC has a 67% chance of appearing, based on NHC's track forecast error statistics over the previous five years. Although newer products have been developed and may be more directly relevant to the hazards (e.g., winds, storm surges), the cone of uncertainty remains the most widely used graphic by the media and the general public (Millet, Carter, et al., 2020).

Notwithstanding its widespread use, this graphic has several shortcomings and is often criticized for leaving out important information (Demuth et al., 2012). Simultaneously, the graphic is overloaded with many different types of information: in addition to forecast uncertainty, the projected track line, and watches and warnings, the graphic also provides a detailed map legend that includes TC classifications (e.g., tropical storm, hurricane, major hurricane). The range of information presented, and the graphic features employed all contribute to visual clutter and information overload (e.g., Eppler & Mengis, 2004).

While the amount and type of information in the overall graphic are problematic, research has also shown that the cone element itself leads to misinterpretation. Some people interpret the cone as a boundary for storm risk (Liu et al., 2015; Ruginski et al., 2016). This leads some people to feel safer in locations just outside the cone limits (Broad et al., 2007; Cox et al., 2013; Wu et al., 2014), or even to believe that there is no imminent risk beyond the cone. By design, however, there is in fact a one-third chance that the center of the storm will not be within the cone at all. In addition, people misinterpret the COU's central black line, not recognizing its inherent uncertainty or the line's function as only depicting the center of the storm. Thus, evidence suggests that people believe that areas along the central line are in greater danger than those in the vicinity of the track line.

CONTACT Barbara Millet bmillet@miami.edu School of Communication, University of Miami, Coral Gables, FL, USA

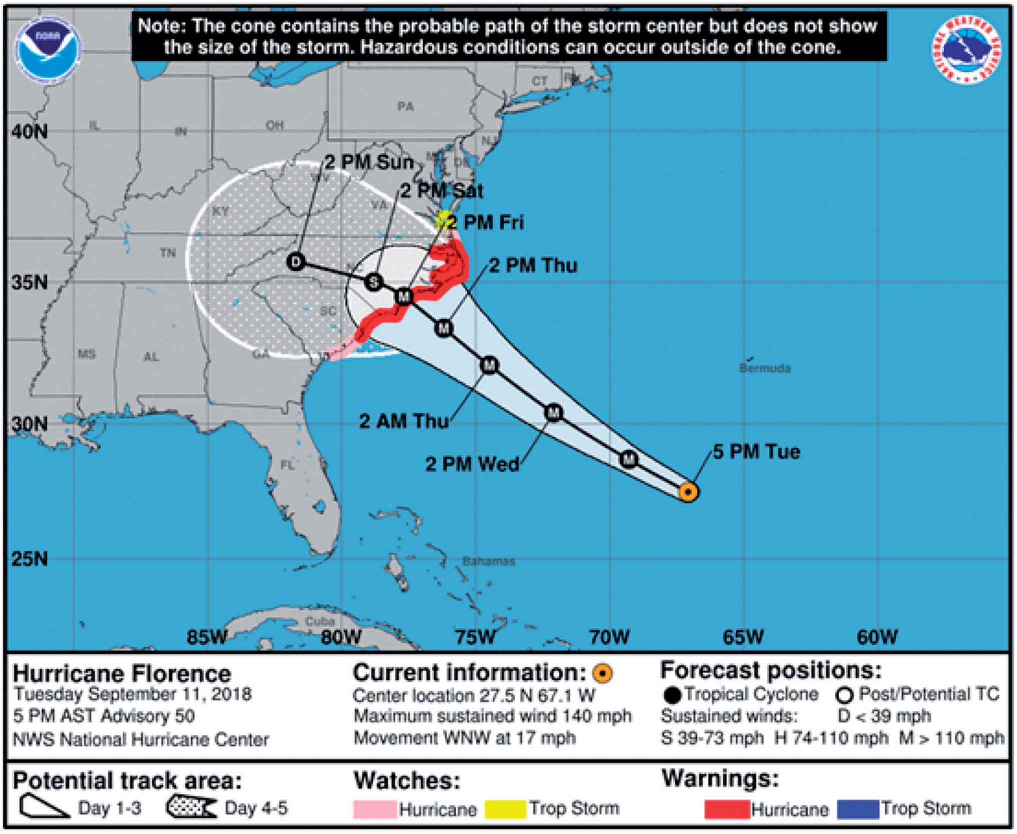
Note: The cone contains the probable path of the storm center but does not show the size of the storm. Hazardous conditions can occur outside of the cone.
40N
35N
30N
25N
85W
80W
75W
70W
65W
60W
2 PM Sun
2 PM Sat
2 PM Fri
2 PM Thu
2 AM Thu
2 PM Wed
5 PM Tue
D
S
M
IL
IN
OH
PA
KY
TN
VA
SC
GA
MS
AL
FL
CT
NJ
Bermuda
Bahamas
Cuba
Hurricane Florence
Tuesday September 11, 2018
5 PM AST Advisory 50
NWS National Hurricane Center
Current information: ⊙
Center location 27.5 N 67.1 W
Maximum sustained wind 140 mph
Movement WNW at 17 mph
Forecast positions:
● Tropical Cyclone ○ Post/Potential TC
Sustained winds: D < 39 mph
S 39-73 mph H 74-110 mph M > 110 mph
Potential track area:
Day 1-3 Day 4-5
Watches:
Hurricane Trop Storm
Warnings:
Hurricane Trop Storm

Figure 1. An example of a hurricane forecast cone typically presented to end users by the National Hurricane Center (NHC 2018; https://www.nhc.noaa.gov/archive/2018/FLORENCE_graphics.php?product=5day_cone_no_line).

Importantly, people also often fail to recognize that the size and intensity of the storm are not represented by the cone element in the graphic. For example, instead of construing the widening shape of the cone as representing more uncertainty as the forecast moves further into the future, users often misread the shape as indicating that the hurricane grows larger over time (Boone et al., 2018; Liu et al., 2015; Padilla et al., 2017; Ruginski et al., 2016). Furthermore, research indicates that people may use a heuristic relating the cone size to storm intensity (Padilla et al., 2017; Ruginski et al., 2016).

To address shortcomings of the COU, researchers have proposed modifications to the cone and developed alternative visualizations to communicate risk. Cox et al. (2013) explored ensemble path visualization focused on showing the uncertainty associated with hurricane predictions. Their alternative approach relied on direct visualization of an ensemble of possible hurricane tracks generated from historical data and current advisory information. Their study compared the ability of non-experts to estimate the spatial distribution of hurricane impact probability based on either their visualization or the NHC's cone of uncertainty. Findings indicated that, in comparison to the COU, their alternative visualization allowed participants to glean more information about path uncertainty; however, overall, the information was more difficult to interpret.

Subsequently, Liu et al. (2015) developed a time-varying ensemble display to provide users with information regarding the predicted state of a storm at a specific time. Their approach relied on estimates of the likelihood of hurricane risk by interpolating simplicial depth values in the path ensemble. In doing so, they developed a time-varying display presenting potential hurricane paths and locations, including a representation of forecast uncertainty and storm characteristics. Although they did not formally evaluate their visualization, they revealed that their graphic also contributed to a misperception that the storm increases in size as the risk region increases.

Other research efforts have explored the impact of visualization type on novice judgments of potential storm damage (Padilla et al., 2017; Ruginski et al., 2016). These studies explored the effects of summary and ensemble displays on interpretations of hurricane uncertainty data. Consistent across the studies, findings indicated that novice users interpreted hurricane size and intensity differently when viewing COUs and ensemble displays.

Alternative hurricane visualization efforts have been focused on improving representations of uncertainty. However, these attempts have not considered the influence of specific design elements on the users' ability to interpret the graphic. Therefore, in this study, we examine how non-expert users interpret visualizations displaying uncertainty in hurricane forecasts, specifically by exploring how visualization design influences the prediction of uncertain spatial trajectories. In a laboratory setting, we compared non-expert

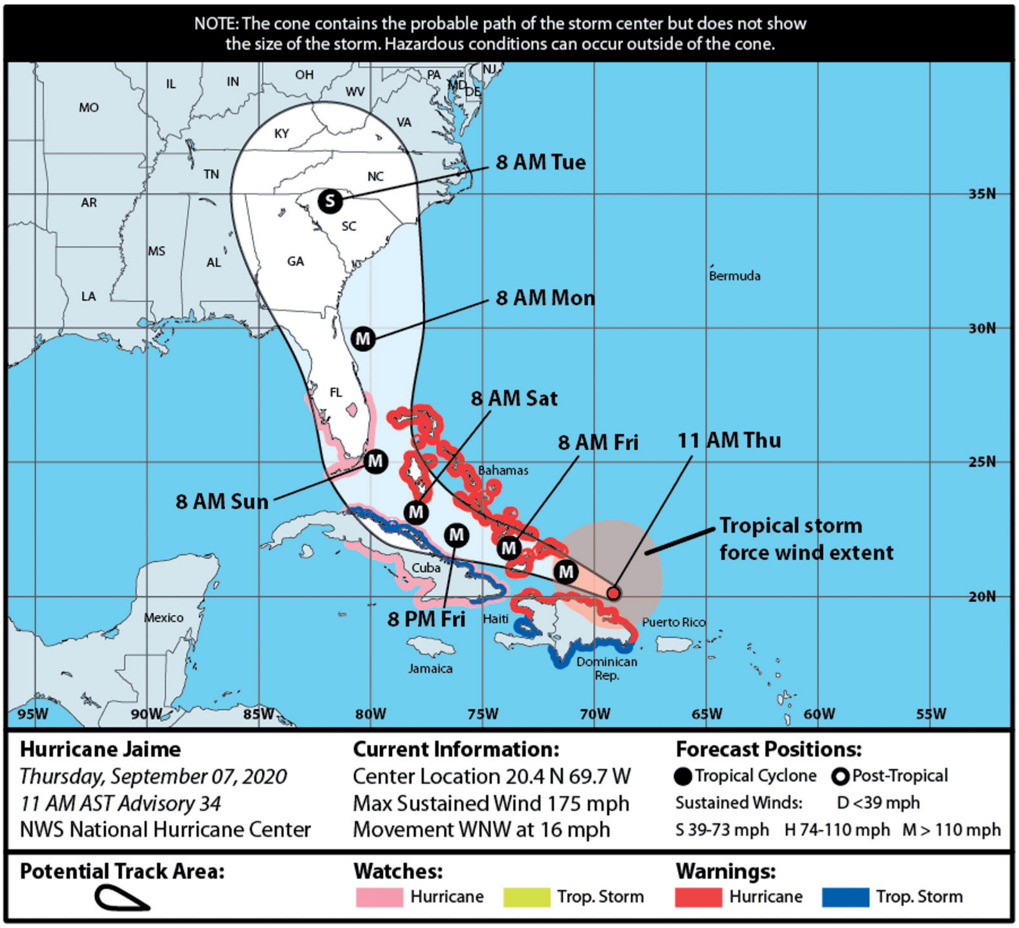
NOTE: The cone contains the probable path of the storm center but does not show the size of the storm. Hazardous conditions can occur outside of the cone.
MO
IL
IN
OH
WV
PA
NJ
MD
DE
VA
KY
TN
NC
SC
AR
MS
AL
GA
LA
FL
8 AM Tue
8 AM Mon
8 AM Sat
8 AM Fri
11 AM Thu
8 AM Sun
8 PM Fri
S
M
Bermuda
Bahamas
Cuba
Mexico
Haiti
Jamaica
Dominican Rep.
Puerto Rico
Tropical storm force wind extent
35N
30N
25N
20N
95W
90W
85W
80W
75W
70W
65W
60W
55W
Hurricane Jaime
Thursday, September 07, 2020
11 AM AST Advisory 34
NWS National Hurricane Center
Current Information:
Center Location 20.4 N 69.7 W
Max Sustained Wind 175 mph
Movement WNW at 16 mph
Forecast Positions:
● Tropical Cyclone ○ Post-Tropical
Sustained Winds: D <39 mph
S 39-73 mph H 74-110 mph M > 110 mph
Potential Track Area:
Watches:
Hurricane
Trop. Storm
Warnings:
Hurricane
Trop. Storm

Figure 2. NHC cone of uncertainty (with minor visual design modifications made by the authors).

responses to the COU product against alternative forecast graphics and explored how elements in the visualization attract attention and influence interpretations of hurricane risk.

2. Methods

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Miami.

2.1. Participants

Forty-three university students recruited via an email list participated in the study. All participants were screened for color vision deficiency. Participants were compensated with a $25 gift card.

2.2. Stimuli

For this study, we produced three visualization designs using Automated Tropical Cyclone Forecasting System (ATCF) data from nine tropical cyclones from 2011 to 2018, overlaid on a map of the United States, and modifying colors, legends, and arrangement of the traditional NHC COU. The traditional COU, which has been operational since 2002, contains a hard boundary whose distance from the forecast point is calculated as the 67th percentile of all NHC forecast errors of tropical cyclone tracks over the previous 5 years. The inside of the cone is colored white with the hard boundary represented by a solid black line. The first design was the existing NHC COU (see Figure 2), with only minor updates to increase color contrast, highlighting the cone over the ocean and landmass.

We developed two alternative versions of the NHC COU to eliminate the hard boundary, a feature known to lead to the common misinterpretation that if one is outside of that boundary, one is safe from the storm. The second design (see Figure 3), our COU Redesign A (RDNA), blurred the boundaries of the cone and used a uniform, diffuse gray shading. In addition, we added a "How to Read the Cone" textual explanation above the map. Colors for watches and warnings were also changed to a sequential color palette ranging from light yellow (for Tropical Storm watches) to dark red (for Hurricane warnings.) The legend at the bottom of the map in this version added contextual explanations of map abbreviations: M for major hurricane, H for hurricane, S for tropical storm, and D for tropical depression.

The third design (see Figure 4), our COU Redesign B (RDNB), used different gradations of the diffuse gray shading. Regions closest to the official NHC forecast position were shaded darker, and regions further were lighter. More

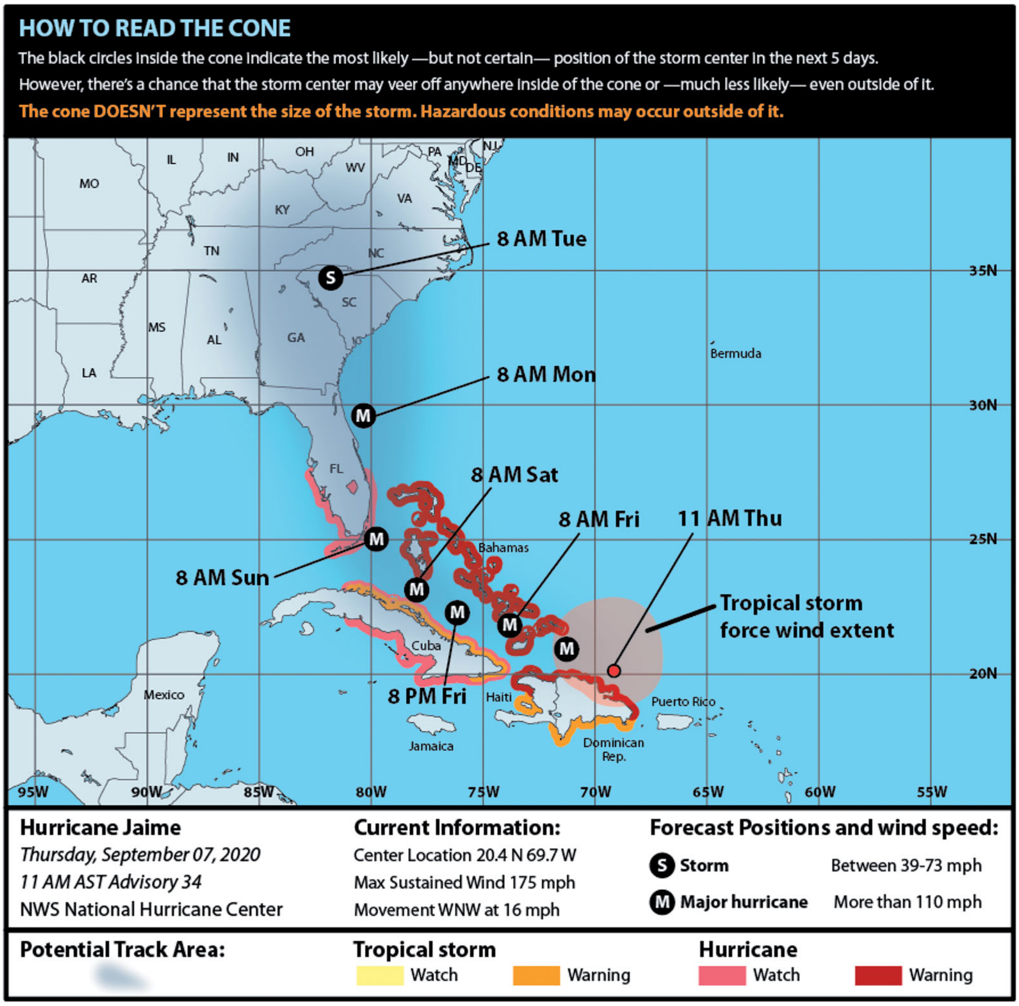
HOW TO READ THE CONE
The black circles inside the cone indicate the most likely —but not certain— position of the storm center in the next 5 days.
However, there's a chance that the storm center may veer off anywhere inside of the cone or —much less likely— even outside of it.
The cone DOESN'T represent the size of the storm. Hazardous conditions may occur outside of it.
MO
IL
IN
OH
WV
PA
NJ
MD
DE
VA
KY
TN
NC
AR
SC
MS
AL
GA
LA
FL
8 AM Tue
8 AM Mon
8 AM Sat
8 AM Fri
11 AM Thu
8 AM Sun
8 PM Fri
Tropical storm force wind extent
Bermuda
Bahamas
Cuba
Haiti
Jamaica
Dominican Rep.
Puerto Rico
Mexico
35N
30N
25N
20N
95W
90W
85W
80W
75W
70W
65W
60W
55W
S
M
Hurricane Jaime
Thursday, September 07, 2020
11 AM AST Advisory 34
NWS National Hurricane Center
Current Information:
Center Location 20.4 N 69.7 W
Max Sustained Wind 175 mph
Movement WNW at 16 mph
Forecast Positions and wind speed:
S Storm
Between 39-73 mph
M Major hurricane
More than 110 mph
Potential Track Area:
Tropical storm
Watch
Warning
Hurricane
Watch
Warning

Figure 3. Redesign A (RDNA) of the NHC cone of uncertainty.

precisely, the 33rd, 50th, and 67th percentiles of the previous 5 years of NHC track forecast errors were computed, using forecast error data provided to our team by the NHC. Circles with a calculated radius for the 33rd percentile were then filled with the darkest diffuse shading, followed by lighter shading for the radius corresponding with the area between the 33rd and 50th percentile, and even lighter shading corresponding to the area between the 50th and 67th percentile. Therefore, the shadings represent a 33% chance that the center of the storm will pass within the innermost cone, a 50% chance that the center of the storm will pass within the innermost or the intermediate shaded cones, and a 67% chance that the center of the storm will pass within an area inside all three cones. This design also modified the NHC COU by depicting separate maps for watches, warnings, and the cone, while also maintaining the changes in the legend and color schemes of the RDNA. In addition, the "How to Read the Map" legend was expanded from the RDNA. Across the designs, fictional names were assigned to each storm in the stimuli, but herein we report the actual names of the storms.

RDNA and RDNB were inspired by a literature review on misinterpretations of hurricane forecast products (Millet, Carter, et al., 2020). As most of the literature on visualization design focuses on the forecast track and its uncertainty, we decided to broaden our analysis to include other important design elements. This shift was confirmed through a user-centered design approach involving eight focus groups conducted in South Florida. This generative research revealed specific design elements, beyond the cone itself, that contribute to misinterpretations of the intended message. The findings from our focus groups directly informed our redesign efforts. For example, participants observed "clutter" in the graphic that, according to them, "was trying to stack" too many layers of information into a single display. This, for instance, led to one of the decisions for RDNB: separating the forecast path from the watches and the warnings, and instead devoting one map to each. Additional findings from the focus groups are provided in Millet, Cairo, et al. (2020) and will be discussed in greater detail in a future article.

2.3. Task

As in Cox et al. (2013), for each trial, participants were asked to estimate the probability that the center of the hurricane would traverse each of eight sectors, corresponding to the cardinal and ordinal points of the compass (N, NE, E,

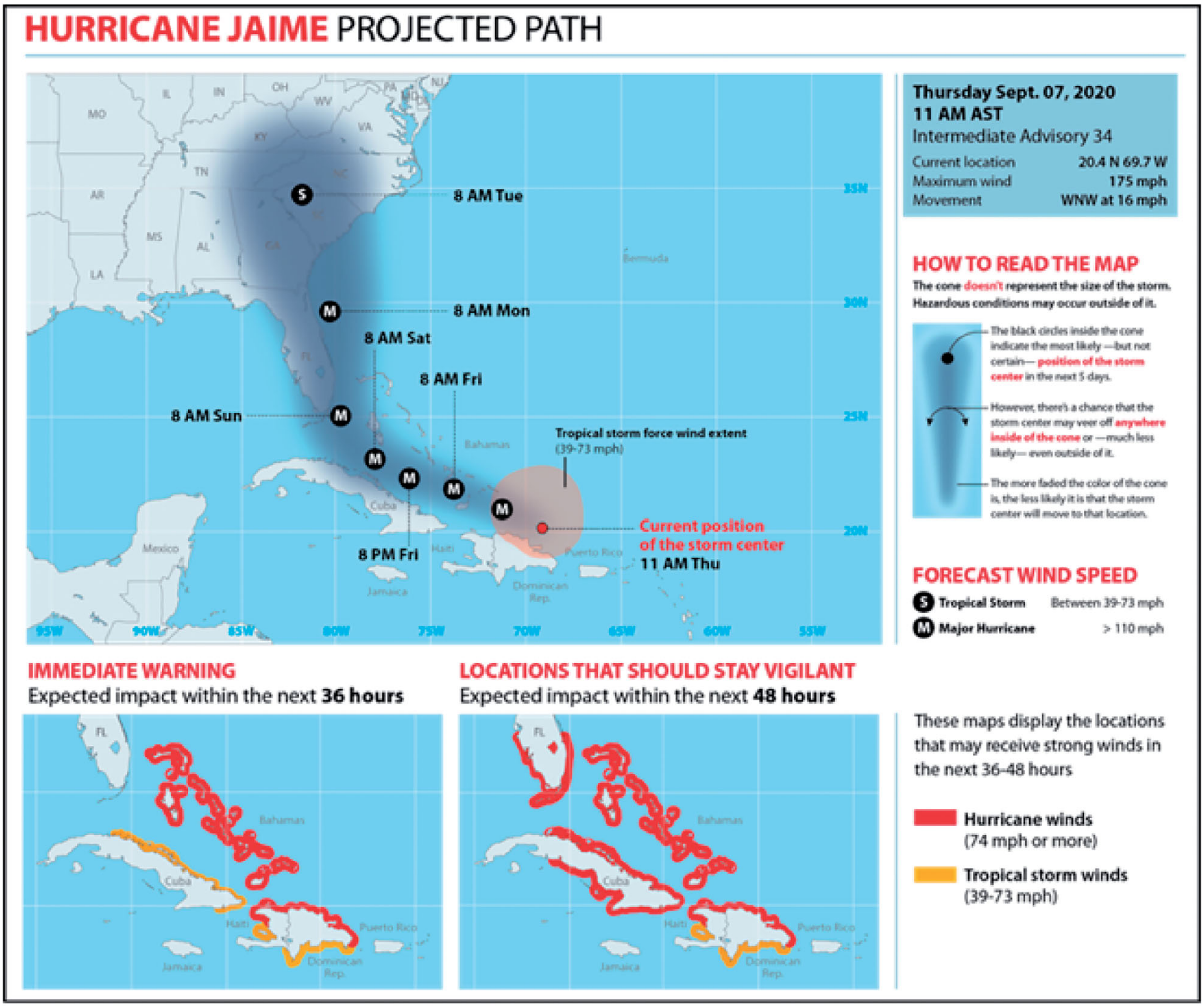
HURRICANE JAIME PROJECTED PATH
Thursday Sept. 07, 2020
11 AM AST
Intermediate Advisory 34
Current location 20.4 N 69.7 W
Maximum wind 175 mph
Movement WNW at 16 mph
MO
IL
IN
OH
PA
NJ
MD
WV
VA
KY
TN
AR
MS
AL
GA
LA
FL
Bermuda
Bahamas
Cuba
Haiti
Puerto Rico
Dominican Rep.
Jamaica
Mexico
8 AM Tue
8 AM Mon
8 AM Sat
8 AM Fri
8 AM Sun
8 PM Fri
Tropical storm force wind extent (39-73 mph)
Current position of the storm center
11 AM Thu
S
M
35N
30N
25N
20N
95W
90W
85W
80W
75W
70W
65W
60W
55W
HOW TO READ THE MAP
The cone doesn't represent the size of the storm. Hazardous conditions may occour outside of it.
The black circles inside the cone indicate the most likely —but not certain— position of the storm center in the next 5 days.
However, there's a chance that the storm center may veer off anywhere inside of the cone or —much less likely— even outside of it.
The more faded the color of the cone is, the less likely it is that the storm center will move to that location.
FORECAST WIND SPEED
S Tropical Storm Between 39-73 mph
M Major Hurricane > 110 mph
IMMEDIATE WARNING
Expected impact within the next 36 hours
LOCATIONS THAT SHOULD STAY VIGILANT
Expected impact within the next 48 hours
These maps display the locations that may receive strong winds in the next 36-48 hours
Hurricane winds (74 mph or more)
Tropical storm winds (39-73 mph)

Figure 4. Redesign B (RDNB) of the NHC cone of uncertainty.

…, NW). These sectors generated 45-degree arcs around the center of the tropical cyclone. The size of the circle encapsulating the sector was held constant across trials and sized to accommodate chip placement. Participants were instructed to place a set of numbered chips in the sectors to indicate their estimate of the probability that the tropical cyclone would exit the circle in the corresponding sector (i.e., strike percentage). The chips ranged in value from 1 to 20 and had a cumulative value of 100. There were two chips valued at 20, four at 10, three at 5, and five 1 s.

2.4. Equipment

The data collection platform was coded using HTML, JavaScript, and D3.js JavaScript library. We used Airtable, a spreadsheet-database hybrid, for data storage. Across all trials, the Airtable database stored participants' ID, along with the sums of chip values assigned to the eight sectors, the name of the storm, the visualization design, and the task time. The data were posted directly through JavaScript via Ajax. The data collection platform was executed from a local folder using a web server for Google Chrome.

Eye tracking, electrodermal activity, and facial expressions were recorded. A Tobii X2-60 eye-tracking system, with a sampling rate of 60 Hz, recorded eye movements while participants completed tasks. Shimmer3 GSR (Galvanic Skin Response) electrodes were attached to two fingers of participants' non-dominant hand. Affectiva's Affdex technology captured participants' facial expressions and provided a classification of emotional states. Affectiva's Affdex and iMotions Facet technology captured real-time frame-by-frame (30 fps) facial expressions from the video stream. All data were collected using the iMotions data integration platform.

2.5. Experimental design and measures

In this study, we examined how interpretations of hurricane forecasts are influenced by design elements. We utilized a mixed design including repeated measures for Tropical Cyclone (9) × Visualization Design (3), resulting in a total of

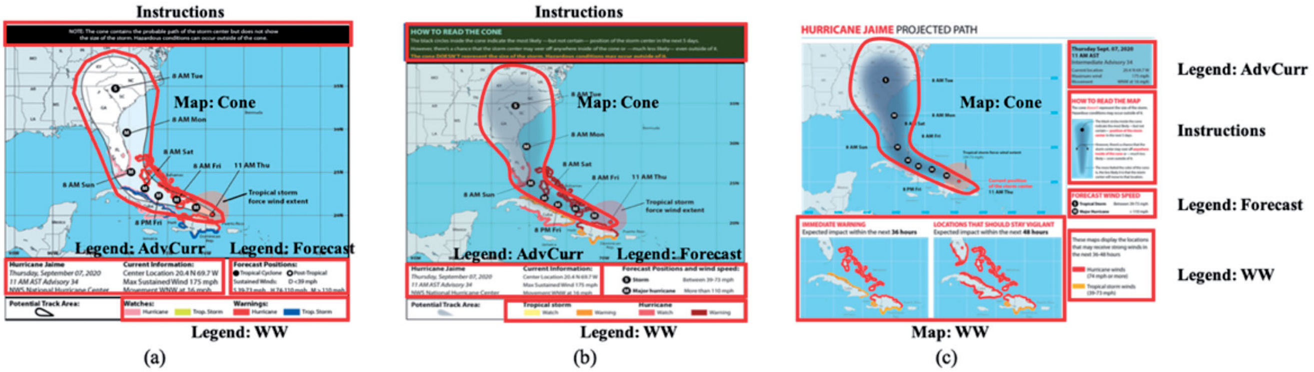

Figure 5. The primary AOIs on each display (a) COU, (b) RDNA, and (c) RDNB. AdvCurr, advisory and current information; WW, watches and warnings.

27 trials per participant. The order of presentation was determined using a diagram-balanced Greco-Latin rectangle (Lewis, 1993) to simultaneously counterbalance the presentation of the visualization method, the hurricane advisory set, and the pairing of the two.

2.5.1. Independent variables

The primary independent variables were two within subject factors, visualization design, and tropical cyclone advisory. The visualization designs were the COU, a modified NHC cone of uncertainty (RDNA), and an alternative visualization (RDNB). The tropical cyclones used were Irene (2011, Advisory 12), Lee (2011, Advisory 2), Isaac (2012, Advisory 7), Sandy (2012, Advisory 19), Matthew (2016, Advisory 27), Harvey (2017, Advisory 15), Irma (2017, Advisory 34), Maria (2017, Advisory 5), and Florence (2018, Advisory 50). The secondary independent variables were two between-subject factors: graphicacy and numeracy. Graphicacy was measured using the Subjective Graph Literacy (SGL) scale developed by Garcia-Retamero et al. (2016). The SGL is a 10-item instrument for estimating perceptions of graph comprehension. Item responses were scored using a 6-point Likert scale, with a score of 1 indicating the least perceived skill. The overall score was calculated by averaging the scores across the 10 items. The validity and reliability (Cronbach's alpha = .87) of SGL was demonstrated previously (Garcia-Retamero et al., 2016). Numeracy was measured using the Subjective Numeracy Scale (SNS) developed by Fagerlin et al. (2007). The SNS is an 8-item instrument that measures participants' preference for the presentation of numerical information and perceptions of their own mathematical ability. Responses are scored using a 6-point Likert scale, with 1 corresponding to the least perceived skill. The overall numeracy score was calculated by averaging the scores across the eight items. We also calculated average scores for the two subscales for preference and mathematical ability. The validity and reliability (Cronbach's alpha = .82) of SNS was demonstrated in an earlier study (Zikmund-Fisher et al., 2007).

2.5.2. Dependent variables

The dependent variables included completion time, comprehension, eye movement, electrodermal activity, valence of emotional expression, mental workload, perceived ease of use (UMUX-Lite), and preference. Completion time was measured, in seconds, from page load until the participant pressed the "Next" button when completing each trial. Comprehension was measured by converting the strike percentage to a discord score, derived by subtracting participant scores from expert opinion and then taking the sum of the absolute value of discord scores across all sectors for each tropical cyclone. This score thus represented the level of disagreement between participants and experts (three meteorologists, all of whom had 15 years or more experience in hurricane forecasting).

Expert sessions were conducted remotely, using the same data collection platform (see section 2.3), but excluded eye movement data and other psychophysiological metrics. Although we gave the storms fictitious data (e.g., names, dates), the experts did recognize a few historical storms. However, the experts were asked to make path estimates based on the data presented and not on their recollections of actual tropical cyclone paths. When the experts recognized the actual storms, they attempted to provide estimates based solely on the advisory information shown and also explicitly pointed out where the actual TC path deviated from the forecast shown. Only the estimates made based on the advisory information presented were recorded. Whether their prior knowledge influenced their estimates is a factor outside of the scope of this study and a possible limitation that can be addressed in future studies.

The eye movement measures were dependent on the assignment of Areas of Interest (AOIs). We defined a set of AOIs on the visualizations corresponding to features containing information essential for completing the task. For analysis purposes, we divided the instructions, map, and legend into more specific bits of information (i.e., Instructions; Map: Cone and Map: Watches & Warnings (RDNB only); Legend: Advisory & Current Information, Legend: Forecast Positions, and Legend: Watches & Warnings), as marked in Figure 5. For each participant, we computed the number and length of time (Fixation Count, Average Fixation Duration, and % Dwell Time) viewing each AOI.

Electrodermal activity was collected to capture arousal response. Electrodermal data was recorded as soon as the participant sat at the workstation to provide reliable baseline values. We also presented a calibration slide—a gray screen—at the start of the experiment. The duration of the baseline stimulus was 60 s. For the analysis, we evaluated the number of galvanic skin response (GSR) peaks that occurred during the trial. GSR only generates information about the arousal dimension; we therefore also included facial expression analysis to capture emotional valence using automated facial expression analysis. The measures for valence indicated the percentage of time an emotional expression fit the classification of positive or negative valence.

The mental workload was assessed with the NASA Task Load Index (NASA-TLX), by calculating the average of the six subscales (mental demand, physical demand, temporal demand, performance, effort, and frustration), where each was scored on a 0–20 scale (Hart & Staveland, 1988). Perceived ease of use was assessed using the UMUX-Lite, a 2-item instrument with a 7-point response scale (Lewis et al., 2013). Preference for visualization was assessed by asking participants to rank the visualizations from 1 (most preferred) to 3 (least).

2.6. Procedure

When participants arrived at the lab, a researcher described the purpose of the research, what to expect during the session, and participants' rights. Informed consent was obtained from each participant. Participation in this experiment was conducted one participant at a time, lasting approximately 60 minutes per session.

Upon starting the study, participants completed a pre-test questionnaire that solicited demographic information and listened to verbal instructions explaining the task and the goal of the experiment. Afterward, participants completed a 9-point calibration eye-tracking exercise. Two electrodes, monitoring electrodermal activity, were attached. To ensure data integrity, distance from the monitor and tracking quality were observed and corrected (if needed) throughout the study.

For each trial, participants were asked to estimate the probability that the center of the TC would traverse each of the eight sectors. We explicitly instructed participants to base their estimates of strike percentage on timepoints at the edge of the sectors. Before the estimation task, participants were presented with the TC visualization and were asked to study the map to determine the TC's path based on the information presented. The preview was identical to the task view, except that the sectors and chips were excluded. The estimation task was self-paced.

When participants finished all trials with a visualization design, the perceived workload was measured using the NASA Task Load (Hart & Staveland, 1988). Participants also rated ease of use for each visualization design using the UMUX-Lite questionnaire (Lewis et al., 2013). After participants completed the questionnaire, they were given a 2-min break, and then the process was repeated with the remaining visualization designs. Finally, participants ranked the visualization methods in order of preference and answered questions about what they liked, disliked, and would change about each visualization.

2.7. Analysis strategy

All analyses were conducted using SPSS version 27.0 (IBM Corp, 2020). A series of mixed-effect models were used (also known as multilevel models; Raudenbush & Bryk, 2002) to account for within-person and between-person variabilities separately, providing an unbiased examination of associations between predictors (visualization design and TC) and outcomes (comprehension, completion time, eye movement metrics). First, we entered visualization design and TC conditions as within-person predictors (at the 1st level; within-person main effect model). Second, pre-test scores of participants' numeracy and graphicacy were entered into the within-person effect model as a between-person covariate (at the 2nd level; within- and between-person main effect model). Only significant results were included for subsequent analyses. Third, we examined the potential two-way interaction effects between within-person predictors (visualization design and TC conditions) and between-person predictors (numeracy and graphicacy). This model is known as a cross-level interaction effect model (with within-person between-person predictors). Interaction effects were examined by entering a product term (e.g., visualization design * numeracy) into the model described in the second step (within- and between-person main effect model). Fourth, we also examined the within-person interaction effects between visualization design and TC on outcomes. Bonferroni correction was applied to adjust the significance level (*p*-value) of the analyses for multiple comparisons of the statistical tests. Restricted Maximum Likelihood (REML) was used to estimate accurately within- and between-person effects (Peugh, 2010). For user preference, we analyzed the rank data using the Friedman test. Last, mixed-design ANOVAs were used to compare subjective metrics capturing mental workload (NASA TLX) and perceived ease of use (UMUX Lite) by visualization design.

3. Results

3.1. Sample characteristics

The 19 male and 24 female participants were between the ages of 19 and 47 ($M = 25.70$, $SD = 6.35$). Most participants (81.4%) had prior experience with the COU graphic, having lived in an area threatened by tropical cyclones.

3.2. Graphicacy and numeracy

Participants had a mean total graphicacy score of 4.67 (SGL: $SD = 0.75$, ranging from 2 to 6). Participants scored lowest in their perceived skill in projecting a future trend from a line chart ($M = 4.19$, $SD = 1.38$, ranging from 1 to 6). Additionally, the participants scored highest in their perceptions of frequency in finding graphical information to be

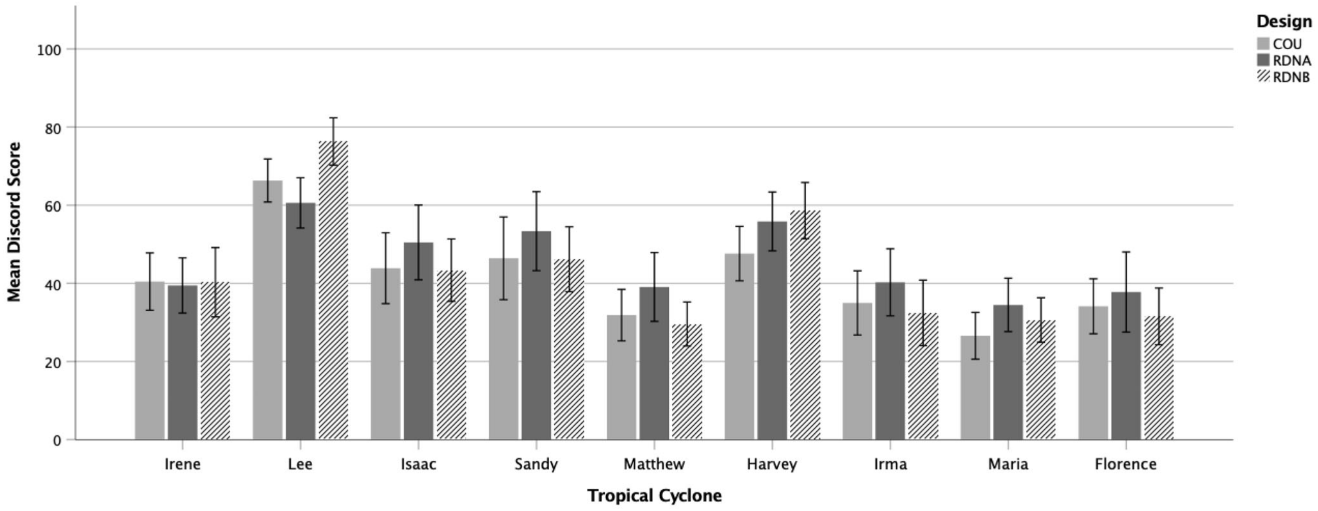

Figure 6. Mean discord score by a tropical cyclone and visualization design, with error bars ±2 SE. A higher discord score corresponds to a lower level of comprehension.

useful ($M = 5.02$, $SD = 0.94$, 3–6). Cronbach's alpha for the SGL scale in our sample was 0.86.

Participants had a mean total numeracy score of 4.56 (SNS: $SD = 0.91$, ranging from 3 to 6). Participants scored lowest in their perceived skill in calculating a 15% tip ($M = 4.02$, $SD = 1.63$, ranging from 1 to 6). In addition, the participants scored highest in their perceptions of frequency in finding numerical information to be useful ($M = 5.16$, $SD = 0.98$, 3–6). Cronbach's alpha for the SNS scale in our sample was 0.78. Graphicacy scores were significantly and positively associated with numeracy scores, $r = .618$, $p < .001$.

3.3. Comprehension

The overall results for average discord score (our proxy for comprehension, where lower discord scores indicated greater comprehension) were COU $M = 41.368$ ($SD = 27.095$), RDNA $M = 45.714$ ($SD = 28.858$), and RDNB $M = 43.209$ ($SD = 28.044$). The overall results by TC for the mean discord score are shown in Figure 6. Associations between visualization design and discord score showed that participants' comprehension varied depending on the visualization design [$F(2, 1108) = 3.438$, $p < .05$]. In particular, the mean discord score for RDNA was significantly greater than for COU ($p < .05$). In post-study interviews, participants indicated that the RDNA cone was difficult to see because the diffused gray shading was too light. Further, we found that participants' comprehension also varied depending on the TC [$F(8, 1108) = 34.62$, $p < .001$]. Specifically, the mean discord scores for TC Harvey and Lee were higher than for others in the set (see Figure 6). Both storms were slow-moving and therefore had more compact cone shapes than the others (see McNoldy, 2022 for information about cone shape).

Even after controlling for the effects of numeracy and graphicacy, associations of visualization design and TC with discord scores remained, with significant associations between visualization design and discord scores [$F(2, 1108) = 3.438$, $p < .05$] and between TC and discord scores [$F(8, 1108) = 34.62$, $p < .001$]. We also found that the associations between numeracy and discord scores were significantly positive [$F(1, 40) = 6.32$, $p < .05$]. However, the associations between graphicacy and discord scores were not significant [$F(1, 40) = 0.00$, $p = .993$].

Results of the cross-level interactions among within-person predictors (visualization design and TC) and the between-person predictor (numeracy) on discord scores indicated no significant interaction effects between visualization design and numeracy on discord scores [$F(2, 1088) = .63$, $p = .52$]. However, an interaction effect between TC and numeracy on discord scores was significant [$F(8, 1088) = 2.640$, $p < .01$]. For interpretation of this significant cross-level interaction effect, subgroup analyses were conducted. Based on a median split ($Mdn = 4.625$), 22 participants were assigned to a higher numeracy group ($M = 5.304$, $SD = 0.417$) and 21 to a lower numeracy group ($M = 3.77$, $SD = 0.565$). We examined how the association between TC and discord scores varied depending on the numeracy group. We found a significant interaction effect between TC and numeracy [$F(8, 1098) = 2.56$, $p < .01$] only for one TC (see Figure 7). That is, when estimating TC Sandy's path, participants who had low levels of numeracy had higher discord scores compared to those who had high levels of numeracy [$M = 60.968$ (low numeracy) *vs.* 36.909 (high numeracy), $p < .001$].

Results of the cross-level interactions among within-person predictors (visualization design and TC) and the between-person predictor (graphicacy) on discord scores indicated no significant interaction effect between visualization design and graphicacy on discord scores [$F(2, 1088) = .47$, $p = .62$]. However, as detected for numeracy, an interaction effect between TC and graphicacy on discord scores was also significant [$F(2, 1088) = 1.97$, $p < .05$]. For interpretation of this significant cross-level interaction effect, subgroup analyses were conducted. Based on a median split ($Mdn = 4.900$), 22 participants were assigned to a higher graphicacy group ($M = 5.25$, $SD = 0.34$) and 21 to a lower graphicacy group ($M = 4.09$, $SD = 0.59$). We examined how the association between TC and discord scores varied depending on the graphicacy group and found a

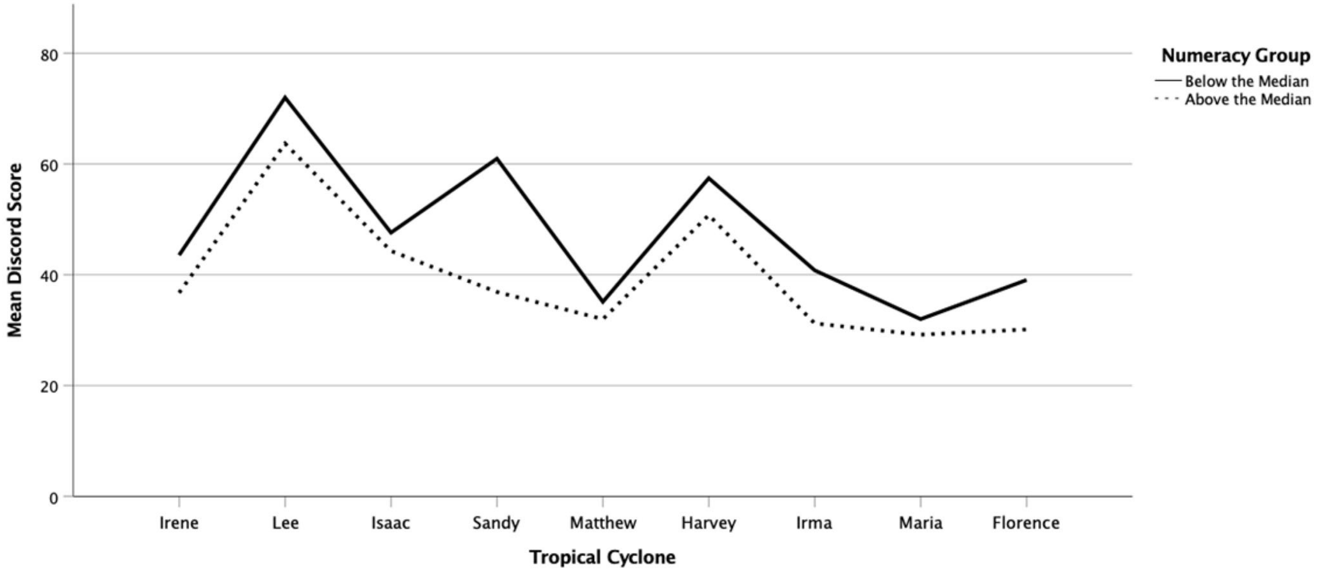

Figure 7. Mean discord score by tropical cyclone and numeracy group. A higher discord score corresponds to a lower level of comprehension.

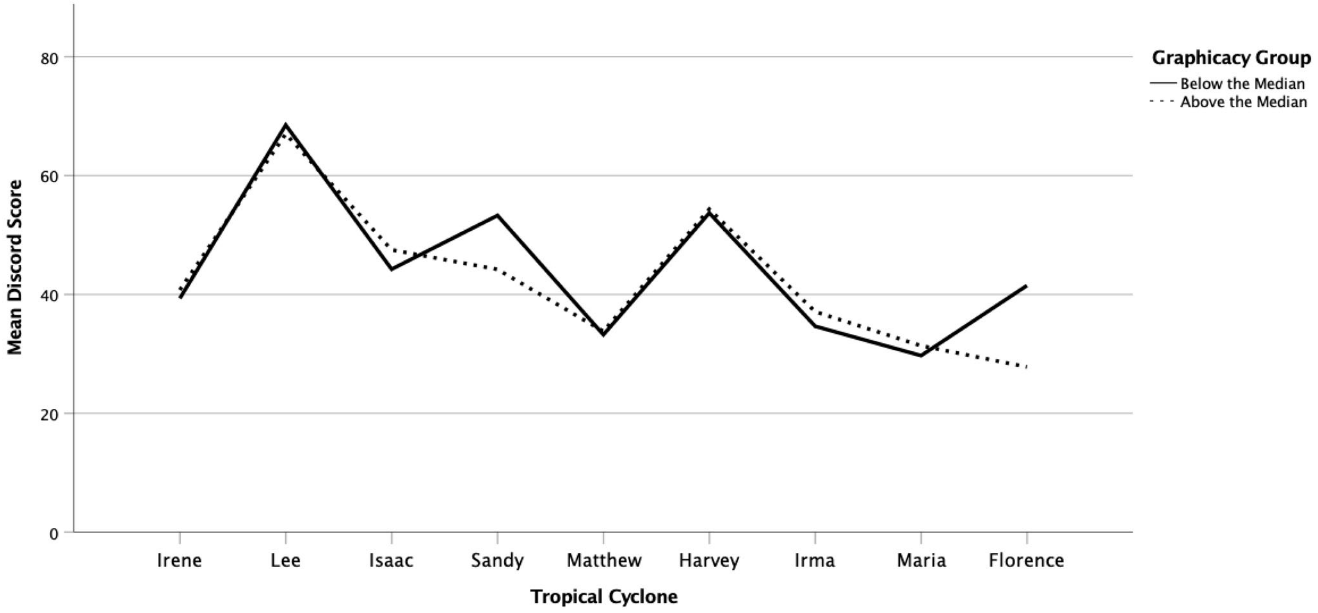

Figure 8. Mean discord score by a tropical cyclone and graphicacy group. A higher discord score corresponds to a lower level of comprehension.

significant interaction effect between TC and graphicacy [$F(8, 1098) = 2.04$, $p < .05$] for only one TC: Florence (see Figure 8). That is, when estimating TC Florence's path, participants who had low levels of graphicacy had higher discord scores compared to those who had high levels of graphicacy [$M = 41.407$ (low graphicacy) *vs.* 27.798 (high graphicacy), $p < .05$]. All other pairwise comparisons were not significant. Furthermore, results indicated non-significant interaction effects between visualization design and TC on discord scores [$F(16, 1092) = 1.54$, $p = .07$].

3.4. Completion time

The overall results for mean completion time, in seconds, by visualization design were COU $M = 62.43$ ($SD = 1.27$), RDNA $M = 64.80$ ($SD = 1.40$), and RDNB $M = 63.39$ ($SD = 1.63$). The overall results for mean completion time by TC are shown in Figure 9. Skewness and kurtosis values for completion time were 2.99 and 15.51, respectively. After log-transformation, skewness and kurtosis for completion time were within the acceptable range (0.98 and 1.44, respectively).

Results indicated no significant associations between visualization design and completion time [$F(2, 1108) = 1.87$, $p = .15$]. However, we did find significant associations between TC and completion time [$F(8, 1108) = 6.46$, $p < .01$]. Even after controlling for the effects of numeracy and graphicacy, associations between visualization design and completion time remained non-significant [$F(2, 1108) = 1.87$, $p = .15$] while associations between TC and completion time remained significant [$F(8, 1108) = 6.46$, $p < .01$]. Results of *post-hoc* analyses indicated that average

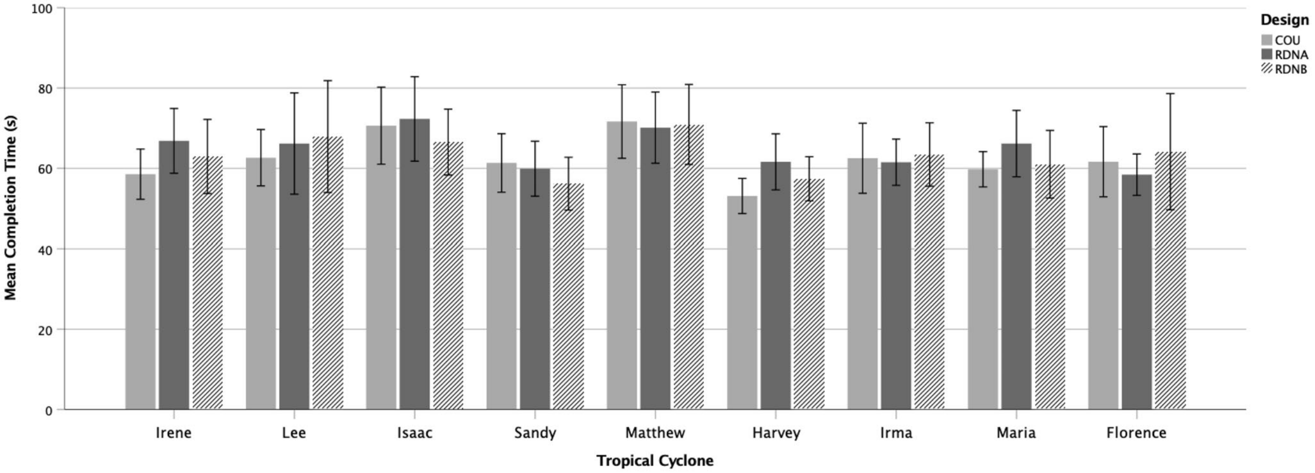

Figure 9. Mean completion time by a tropical cyclone and visualization design, with error bars ±2 SE.

completion times for TCs Matthew and Isaac were significantly higher ($M = 70.91$ for Matthew; $M = 69.82$ for Isaac) than for TCs Florence, Harvey, and Sandy ($M = 61.41$ for Florence; $M = 57.38$ for Harvey; $M = 59.14$ for Sandy; all at $p < .05$). Further, the average completion time of TC Matthew was significantly higher compared to TC Irene ($M = 62.79$; $p < .05$). However, the associations between numeracy and graphicacy themselves on completion time were not significant [$F(1, 40) = .24$, $p = .63$ for associations between numeracy and completion time; $F(1, 40) = .57$, $p = .451$ for associations between graphicacy and completion time].

For cross-level interactions, results of within-person predictors (visualization design and TC) and the between-person predictor (numeracy) on completion time indicated a significant interaction effect between visualization design and numeracy on completion time [$F(2, 1088) = 3.945$, $p < .05$]. However, the interaction effect between TC and numeracy on completion time was not significant [$F(8, 1088) = .257$, $p = .979$]. For interpretation of the significant cross-level interaction effect, subgroup analyses were conducted. Results indicated that participants in the lower numeracy group completed path estimates more quickly when using COU ($M = 59.35$) than when using RDNA ($M = 64.05$; $p < .05$). However, this pattern was not detected for participants in the higher numeracy group ($p > .05$) who had similar completion times for both the COU ($M = 65.38$) and RDNA ($M = 65.50$). Results also indicated a significant interaction effect between visualization design and graphicacy on completion time [$F(2, 1088) = 12.79$, $p < .01$]. However, the interaction effect between TC and graphicacy on completion time was not significant [$F(2, 1088) = .25$, $p = .98$]. We examined how the association between visualization design and completion time varied depending on the graphicacy group and found that participants in the lower graphicacy group completed path estimates more quickly when using COU ($M = 60.60$) than when using RDNA ($M = 70.17$; $p < .001$) and RDNB ($M = 67.34$; $p < .05$). However, as in the numeracy results, this pattern did not hold true for participants in the higher graphicacy group. Furthermore, results indicated no significant interaction effect between visualization design and TC on completion time [$F(16, 1092) = .64$, $p = .85$].

3.5. Eye movement

We investigated where participants allocated attention to the visualization designs when making path estimates. We identified the number of times (fixation count), the length of time (fixation duration), and the percent of total time (% dwell time) per AOI. The eye movement data did not meet normality assumptions and was log-transformed for statistical analyses (as in Hohenstein et al., 2017). For fixation count, skewness and kurtosis values were 7.313 and 95.466, respectively. After log-transformation, skewness and kurtosis were within the acceptable range (1.471 and 1.220, respectively). For fixation duration, skewness and kurtosis values were 2.214 and 14.978, respectively. After log-transformation, skewness and kurtosis were within the acceptable range (0.376 and 1.818, respectively). Last, for % dwell time, skewness and kurtosis values were 3.729 and 16.943, respectively. After log-transformation, skewness and kurtosis were within the acceptable range (0.268 and 1.120, respectively).

3.5.1. Fixation count

Fixation count, measured as the number of fixations on each AOI across trials, varied by visualization design and TC. Figure 10 shows the means for fixation count for the different AOIs. Across the visualization designs and TCs, participants had significantly higher numbers of fixations on the cone element (Map: Cone) compared to all other AOIs (see Table 1). For example, the mean of the Map: Cone AOI (17.147) was higher than the mean of the Key: Watches & Warnings AOI ($M = 0.457$, $t = 1.752$, $p < .001$). For COU and RDNA, the Instructions AOI received the second highest fixation count. This is likely because the instructions are at the top of the visualization design, aligning with top-down viewing. RDNB had a significantly lower fixation count for this AOI than COU ($t = -.533$, $p < .001$) and RDNA ($t = -.683$, $p < .001$), likely because RDNB presented the instructions mid-page, on a secondary panel. Another

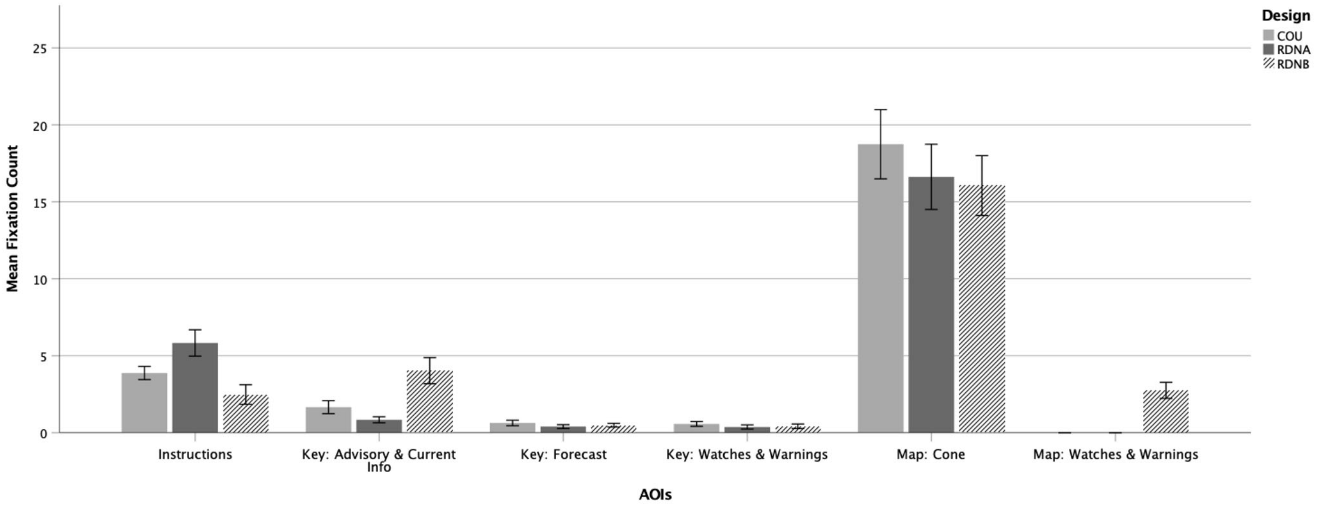

Figure 10. Mean fixation counts for each AOIs by visualization design type, with error bars ±2 SE.

Table 1. Fixation count pairwise comparisons for the Map: Cone AOI.

	AOIs	Pairwise comparisons
Map: Cone	Instructions	$t = 1.022$, $p < .001$
	Key: Advisory & Current Information	$t = 1.432$, $p < .001$
	Key: Forecast	$t = 1.727$, $p < .001$
	Key: Watches & Warnings	$t = 1.752$, $p < .001$

contributing factor may be the improved clarity of the instructions in RDNB, necessitating fewer revisits to achieve comprehension. The topmost AOI for RDNB, the Key: Advisory & Current Information AOI, having the same information across all visualization designs, had the second highest fixation counts. RDNB had significantly greater fixation counts for this AOI than COU ($t = .320$, $p < .001$) and RDNA ($t = .432$, $p < .001$). The Map: Watches and Warnings AOI, in RDNB (the only visualization design to explicitly separate watches and warnings), had the third highest fixation counts ($M = 2.755$, $SD = 7.271$), indicating that participants relied on watch and warning cues to assist in the path estimation task.

3.5.2. Fixation duration

Fixation duration, representing how long participants looked at a specific AOI, measured as the average duration for each AOI across trials in milliseconds (ms), varied by visualization design and TC. Figure 11 shows the means for fixation duration for each AOI of the different visualization designs. Across the visualization designs and TCs, participants had significantly longer fixation duration, on average, on the cone element (Map: Cone) compared to all other AOIs (see Table 2). For COU and RDNA, the Instructions AOI received the second longest fixation duration. Similar to findings from fixation count, this is likely because the instructions are at the top of the visualization design aligning with top-down viewing. RDNB had a significantly shorter fixation duration for the Instructions AOI than COU ($t = -1.483$, $p < .001$) and RDNA ($t = -1.688$, $p < .001$), even though the instructions in RDNB were much longer than in the other designs. For RDNB, participants had the second longest fixation duration on the Key: Advisory & Current Information AOI. For this AOI, RDNB received a significantly greater fixation duration than COU ($t = .811$, $p < .001$) and RDNA ($t = .970$, $p < .001$). The Map: Watches and Warnings AOI in RDNB had the third longest fixation duration. Additionally, there were no significant differences, by visualization design, in average fixation duration for the Key: Forecast and for the Key: Watches & Warnings AOIs (see Table 3), indicating that design improvements, such as using a sequential-qualitative color palette (as in the Key: Watches & Warnings AOI) or providing contextual legends of expected wind force (as in the Key: Forecast AOI), did not alter viewing time.

3.5.3. Dwell time

Dwell time represents the amount of time participants fixated on a specific AOI, as a percentage of total time spent on the trial. Figure 12 shows the mean dwell time for the different AOIs of the visualization designs. Across the visualization designs and TCs, participants had significantly longer dwell time, on average, on the cone element (Map: Cone) compared to all other AOIs (see Table 4). However, participants had a longer dwell time on the Map: Cone AOI when interacting with the COU design than with RDNA ($t = .134$, $p < .001$) or RDNB ($t = .135$, $p < .001$). The Instructions AOI received the second longest dwell time across visualization designs. RDNB had a significantly shorter dwell time for the Instructions AOI than either COU ($t = -.313$, $p < .001$) or RDNA ($t = -.453$, $p < .001$), while COU had a significantly shorter dwell time than RDNA ($t = -.140$, $p < .001$). The increased message length in RDNA, compared to COU, may have contributed to longer dwell time, although an even greater message length did not have the same effect for RDNB. Participants had the second longest dwell time on the Advisory & Current Information (Key) when using RDNB. RDNB had a significantly longer dwell time for this AOI than COU ($t = .218$, $p < .001$) or RDNA ($t = .308$, $p < .001$). The watch and warnings maps, in RDNB, had the third longest dwell time. There were no statistically significant differences, by

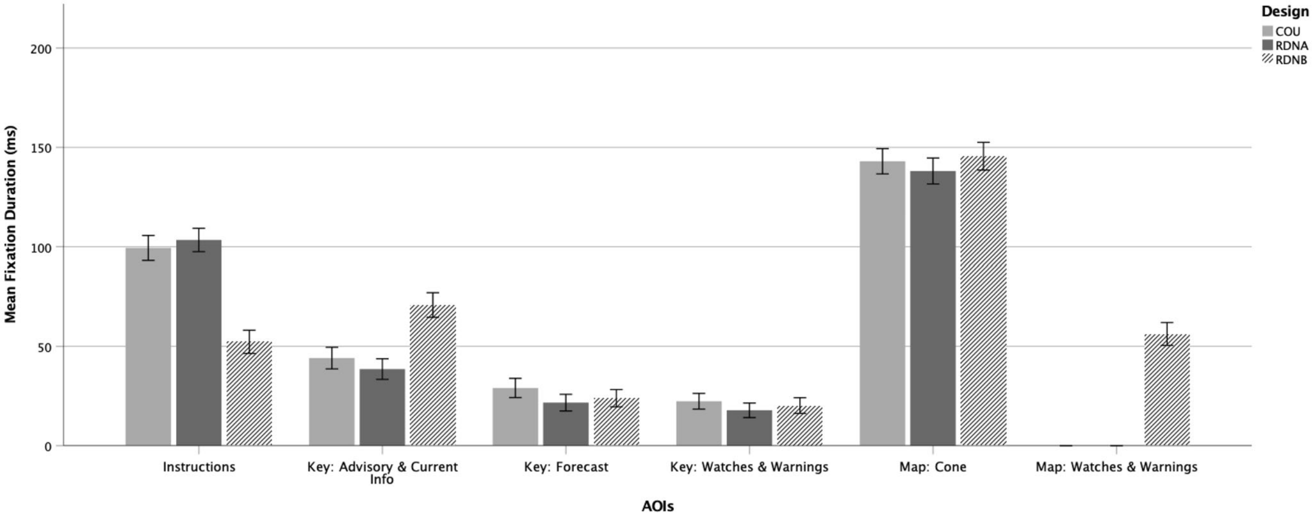

Figure 11. Mean fixation duration for each AOIs, across trials, by visualization design, with error bars ±2 SE.

Table 2. Average fixation duration pairwise comparisons for the Map: Cone AOI.

	AOIs	Pairwise comparisons
Map: Cone	Instructions	$t = 1.404$, $p < .001$
	Key: Advisory & Current Information	$t = 2.472$, $p < .001$
	Key: Forecast	$t = 3.308$, $p < .001$
	Key: Watches & Warnings	$t = 3.416$, $p < .001$

Table 3. Average fixation duration pairwise comparisons for the key: forecast and key: watches and warnings AOIs by visualization design.

	Visualization designs	Pairwise comparisons
Key: Forecast	COU-RDNA	$t = .195$, $p > .05$
	COU-RDNB	$t = .146$, $p > .05$
	RDNA-RDNB	$t = -.050$, $p > .05$
Key: Watches & Warnings	COU-RDNA	$t = .152$, $p > .05$
	COU-RDNB	$t = .098$, $p > .05$
	RDNA-RDNB	$t = -.054$, $p > .05$

visualization design, in dwell time for the Key: Forecast and Key: Watches & Warnings AOIs (see Table 5) similar to the patterns identified previously.

3.6. Workload

Data were analyzed to examine associations between visualization design, numeracy, graphicacy, and overall workload. The overall mean workload score for the forecast graphic designs was 32.8 for the COU, 32.6 for RDNA, and 34.1 for RDNB. Results from a mixed-design ANOVA indicated no statistically significant differences in overall workload scores by visualization design [$F(2, 78) = .580$, $p = .562$], graphicacy [$F(1, 39) = 3.010$, $p = .091$], numeracy [$F(1, 39) = 0.837$, $p = .366$], or their interactions.

3.7. Ease of use ratings

The ease of use ratings for each visualization design was calculated by computing the UMUX-LITE score. The overall mean score for the visualization designs was 69 for the COU, 65 for RDNA, and 67 for RDNB. Results from a mixed-design ANOVA indicated no statistically significant differences in UMUX-Lite scores by design [$F(2, 78) = 1.73$, $p = .18$], graphicacy [$F(1, 39) = 0.094$, $p = .76$], numeracy [$F(1, 39) = 0.87$, $p = .36$], or their interactions.

3.8. Preference

The mean ranks for visualization designs were 1.81 for the COU design, 2.07 for RDNA, and 2.12 for RDNB (a lower mean rank is better-closer to first place). The COU design received the most first-place votes (19/43, 44%). When selecting the COU as their preference, participants indicated that familiarity with the graphic played an important role. In fact, of the 19 participants who ranked the COU first, 58% favored RDNA as their second choice, suggesting a preference for the visualization design which more closely aligned to the COU. Analysis with a Friedman test, however, showed no significant effect of visualization design on preference [$X^2(2) = 2.279$, $p > .05$].

3.9. Post-session interviews

Post-task interviews revealed additional factors, beyond familiarity, that influenced participants' visualization design preference and interpretation. Cone design treatment was a primary reason for top choice selection across the three visualization designs. Other factors included treatment for watches and warnings, color scale, layout, and legend (Millet, Cairo, et al., 2020).

3.9.1. The cone

Of the 19 participants who selected the COU as their most preferred, 18 indicated that their preference was because the cone design's high visual saliency drew their attention to the perceived focal point of the graphic. However, of the five participants providing interpretations, all provided erroneous interpretations of what the cone is meant to convey, such as indicating that the black border provides a boundary for the location and impact of the storm and contains all possible storm paths. Although the cone design was the top reason

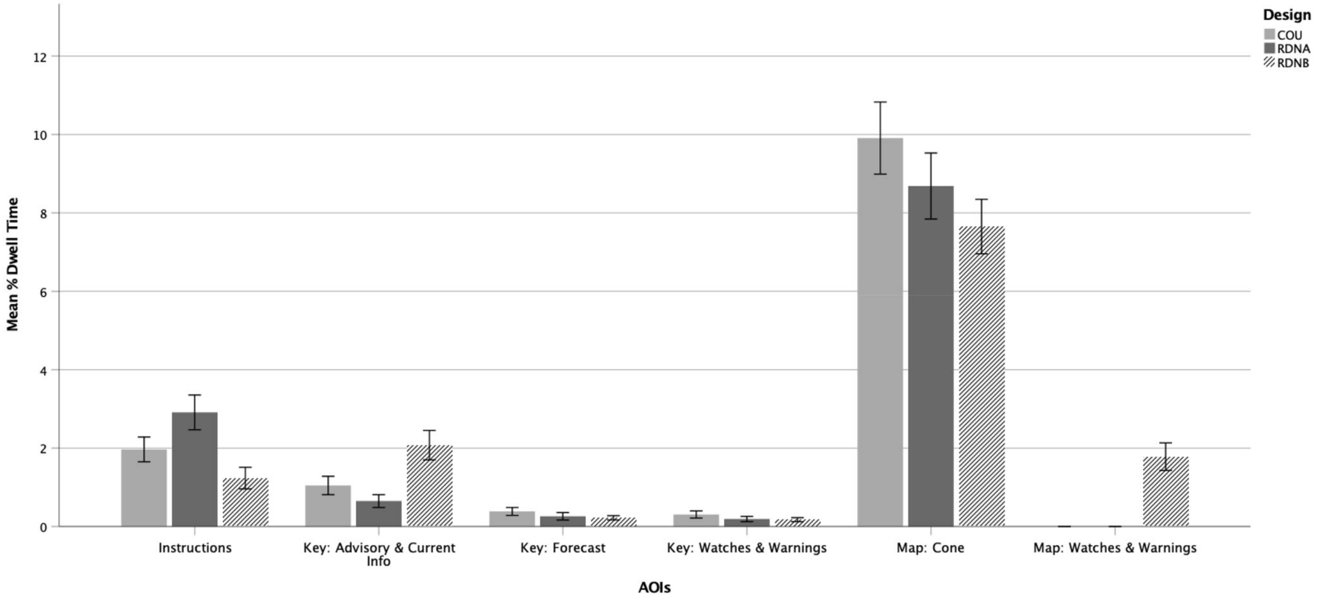

Figure 12. Mean dwell time for each AOIs by visualization design, with error bars ±2 SE.

Table 4. Average dwell time pairwise comparisons for the Map: Cone AOI.

	AOIs	Pairwise comparisons
Map: Cone	Instructions	$t = .960$, $p < .001$
	Key: Advisory & Current Information	$t = 1.205$, $p < .001$
	Key: Forecast	$t = 1.437$, $p < .001$
	Key: Watches & Warnings	$t = 1.461$, $p < .001$

Table 5. Average dwell time pairwise comparisons for the key: forecast and key: watches and warnings AOIs by visualization design.

	Visualization designs	Pairwise comparisons
Key: Forecast	COU-RDNA	$t = .056$, $p > .05$
	COU-RDNB	$t = .045$, $p > .05$
	RDNA-RDNB	$t = -.011$, $p > .05$
Key: Watches & Warnings	COU-RDNA	$t = .044$, $p > .05$
	COU-RDNB	$t = .037$, $p > .05$
	RDNA-RDNB	$t = -.007$, $p > .05$

for participants' first choice of the COU, cone design was also important to those who selected RDNA (11) and RDNB (13), albeit to a lesser extent (63 and 46%, respectively). Participant responses included both correct and incorrect interpretations. The incorrect interpretations of the shaded ("fuzzy") cone, used in RDNA and RDNB, included that a fuzzy boundary implies that the warning is less urgent and indicates that the further you are from the center of the cone, the less likely you are to feel storm effects. Correct interpretations noted that the shading depicts the probability that the storm can pass outside of the cone. For the RDNB design, in particular, a few participants indicated that the shaded cone with the gradient did a better job of conveying the uncertainty in the forecast and eliminated "a false sense of security."

3.9.2. Watches and warnings

Of the 13 participants who selected the RDNB visualization as their most preferred, seven explained that presenting the watches and warnings separately influenced their preference. For some, the separation of information better highlighted coastal areas under threat. However, more participants (10) stated directly that overlaid watches and warnings were critical for storm path estimation. This feedback aligns with eye movement data, suggesting that participants believe that watch and warning information improves their understanding of the forecast path.

3.9.3. Color scale

Participants also found the use of a gradient color palette problematic when conveying categories of tropical cyclones (i.e., tropical storm *vs.* hurricane) and alerts (i.e., watches and warnings). Specifically, they disliked the use of the sequential, warm palette ranging from yellow to red in the RDNA visualization. Few participants (5) observed that the color palette did not provide sufficient discrimination between a tropical storm warning and a hurricane watch, making it difficult to see the differences. As a result, more participants (7) preferred the COU's use of distinct colors to convey categories of tropical cyclones and alerts. Participants did not identify the color scale used in RDNB as problematic, possibly because different colors were used to convey the tropical cyclone categories, while separate maps conveyed each alert type.

3.9.4. Layout

Eleven participants expressed dissatisfaction with RDNB, indicating that the visualization had higher search demands: information hierarchy was inadequate; related information was not grouped; and more processing was required for the additional data elements. The participants felt the layout was unclear, with design elements sequenced without apparent reason, making the message difficult to discern. Furthermore, the arrangement of the elements in RDNB placed related items distally. For example, the "how to read map" component was placed in between the advisory information and the legend for the forecast wind speed.

Participants explained that because the advisory information and forecast legend were related, they should be grouped together. Five participants also noted that RDNB had twice as many design elements to process than the other two visualizations, increasing the complexity of the visual representation of the information.

3.9.5. Legend

A common criticism of the COU was that it did not provide precise labeling of storm categories. The map has icons representing the category of tropical cyclones (e.g., D, S, H, M), but the legend does not provide a textual description of the meaning of the letters (e.g., M = Major Hurricane). Regarding RDNA and RDNB, participants appreciated that the legend provided matching and contextually relevant information about tropical cyclone intensity.

4. Discussion

The cone of uncertainty is one of the most widely used hurricane forecast graphics, but research shows that the general public often misinterprets the information conveyed, interfering with timely preparation and overall decision-making. To better communicate hurricane risk information, alternative visualizations should be considered. In this study, we explored how visualization design elements and users' numerical and graphical literacy influence interpretations of hurricane forecast information.

We found that participants had significantly better comprehension scores when using the COU graphic compared with RDNA. RDNA differed from the COU in four ways: (1) colors for watches and warnings were changed to a sequential-qualitative color palette to better align with representations of tiered, binary variables, (2) contextual explanations of map abbreviations were added to the legend at the bottom of the map to provide situationally relevant information, (3) expanded instructions were placed at the top of the map to provide additional informational, and (4) diffused gray shading with a fuzzy boundary was used to represent the forecast and associated uncertainty to alleviate the common misinterpretation of safety in areas outside of the COU's cone. However, eye movement data showed that participants rarely attended to the legend, expanded instructions in RDNA resulted in longer dwell times, which is to be expected, and in post-study interviews, participants indicated that the RDNA cone was difficult to see because the diffused gray shading was too light. The use of transparency and fuzzy boundaries may have resulted in the cone blending in with the background map, affecting user comprehension. Therefore, the use of transparency and boundary fuzziness when encoding data uncertainty in hurricane forecast graphics is not recommended. However, our findings do not provide direct evidence of whether sequential-qualitative color palettes, contextual legends, and expanded instructions contribute to the overall understanding of hurricane forecast products. Further investigations are needed to determine if a sequential-qualitative color scheme, in particular, representing the two alert levels by varying color hue (watches and warnings) and the two intensities of tropical cyclones by varying color value (tropical storm, hurricane), is effective for weather forecast products. Furthermore, we found that participants' comprehension of hurricane forecast information depended less on visualization design and more on subjective numeracy, subjective graphicacy, and tropical cyclone characteristics that influence cone shapes, such as translational speed and changes in heading.

We found two significant interaction effects. One was an interaction effect between TC and numeracy. That is, when estimating tropical cyclone Sandy's path, participants who had low levels of numeracy had lower comprehension scores than those who had high levels of numeracy. Hurricane Sandy differed from the other tropical cyclones in the set, in that its cone depicted an S-pattern, the watches and warnings issued were outside the cone's boundary (advisory 19), and it began as a tropical cyclone and then transitioned to a post-tropical cyclone, as conveyed in the legend, just before landfall. Sandy's S-shaped cone reflected its shift westward toward land rather than the more typical turn to the northeast over the north Atlantic. Because the cone did not overlap with the watches and warnings, we observed that participants incorporated watch and warning depictions into their path estimates rather than relying on the specific path attributes. Furthermore, the term post-tropical may have confused some participants given that there is no explanation of what the term means in the graphic. Participants may not have understood that a post-tropical cyclone can still produce sustained winds of hurricane and tropical storm force. The second significant interaction effect was between TC and graphicacy on comprehension scores. That is, when estimating tropical cyclone Florence's path, participants who had low levels of graphicacy had lower comprehension scores compared to those who had high levels of graphicacy. Hurricane Florence had the greatest rate of change in translational speed of all tropical cyclones in the set, decreasing from a peak of 14.9 to 2.6 kt in just 3 days. These findings further support the evidence that cone shape impacts comprehension. In addition to a cone shape, however, there may be other characteristics that also impact comprehension that remain unknown at this time. Overall, our results indicate that numeracy, graphicacy, and tropical cyclone characteristics, in combination, influence interpretations of the hurricane forecast track.

We observed that familiarity is a driver for preference, confirming previous studies (Bornstein, 1989) reporting that repeated exposure to a stimulus increases the likelihood of participants choosing that stimulus over others. Regarding completion time, we found that participants with lower subjective numeracy and graphicacy took longer to complete path estimation tasks with the alternative designs than with the one they are familiar with (i.e., the COU), but this extra processing was not needed for participants with higher subjective numeracy and graphicacy scores. These findings suggest that users with higher numeracy and graphicacy may be more efficient when reading alternative displays.

Across the visualization designs and tropical cyclones, regardless of the design modifications, participants had significantly higher numbers of fixations, longer fixation durations, and longer dwell times on the cone element compared to all other AOIs. We also observed that elements placed at the top of the display received greater visual attention, in that the instructions, which are placed at the top, for COU and RDNA, had the second highest numbers of fixations, fixation durations, and dwell times. When the instructions were placed in another location, as in RDNB, the AOI received less visual attention. Eye movement data offered evidence that our design modifications did not alter visual attention patterns. Interestingly, how long and how many times participants looked at different aspects of the visualization did not influence their estimates of the forecast path.

A significant design change for RDNB was separating watch and warning information from the forecast track to alleviate information clutter and facilitate understanding. However, this design modification did not improve understanding. We also found that participants relied on watch and warning information when making decisions about forecast path as evidenced by the eye movement data, suggesting a misinterpretation, as watches and warnings do not convey track information, only wind risk, and only for the coast. This misinterpretation usually does not affect users' judgment, as the information tends to be congruent, but it does suggest that participants use impact-based, rather than just track-based information, in making predictions about hurricane track. We further found that comprehension of the forecast track did not improve in response to changes in cone style (e.g., bounded *vs.* fuzzy border), use of a sequential-qualitative color palette, expanded instructions, and inclusion of contextual legends. Our findings, therefore, do not support efforts relying solely on visual design modifications and suggest that the users' interpretive difficulties are inherent to the display type and even the information offered.

5. Conclusion

Hurricane forecast graphics have the challenging task of communicating both complex information about hazards and spatio-temporal uncertainty. Unfortunately, the general public often misinterprets forecast graphics. Providing effective visualizations is critical for successful messaging of hurricane risks. This study examined the influence of design modifications on comprehension and completion time when participants estimated the forecast path of nine tropical cyclones. We supplemented this analysis with eye-tracking measures to examine the areas of the visualizations with the greatest fixation counts, longest fixation durations, and greatest dwell times. Overall, the design modifications appeared to have limited influence on comprehension and response time. Furthermore, we found that non-experts with lower levels of graphicacy or numeracy may have greater difficulty in understanding path uncertainty in forecasts depicting changes in translational speed or multiple changes in heading. We also found that visual attention in this complex decision task was mostly focused on the more salient information (e.g., the cone), while other information was often ignored. Although the participants preferred the COU, familiarity did not mean that users could correctly interpret the information provided in that graphic. Marginally alternative designs were not found to be more effective. Our study aligns with the literature showing that users have trouble understanding forecast uncertainty and adds to this literature in three important ways: (1) we found that confusion extends beyond the cone, and applies to other design elements; (2) we show, however, that users do rely on other salient aspects of the graphic in addition to the cone; and (3) we found that user characteristics, such as graphicacy and numeracy, also contribute to misinterpretations for specific storm characteristics. This study provides evidence that future design efforts should be focused on completely reimagined visualizations for communicating tropical cyclone forecast paths. Furthermore, based on the findings from this study and our previous work delineating non-experts' unmet needs for concrete, actionable information, we recommend that future design efforts be focused on visualizations that emphasize hazards and risk, rather than just the possible path of the storm center.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the University of Miami Laboratory for Integrative Knowledge Program (Award PG012061).

ORCID

Barbara Millet http://orcid.org/0000-0002-2618-2186
Sharanya J. Majumdar http://orcid.org/0000-0001-8909-0927
Brian D. McNoldy http://orcid.org/0000-0003-0217-1025
Scotney D. Evans http://orcid.org/0000-0003-0897-0725

References

Boone, A. P., Gunalp, P., & Hegarty, M. (2018). Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *Journal of Experimental Psychology. Applied*, *24*(3), 275–295. https://doi.org/10.1037/xap0000166

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*(2), 265–289. https://doi.org/10.1037/0033-2909.106.2.265

Broad, K., Leiserowitz, A., Weinkle, J., & Steketee, M. (2007). Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, *88*(5), 651–668. https://doi.org/10.1175/BAMS-88-5-651

Cox, J., House, D., & Lindell, M. (2013). Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification*, *3*(2), 143–156. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2012003966

Dash, N., & Gladwin, H. (2007). Evacuation decision making and behavioral responses: Individual and household. *Natural Hazards Review*, *8*(3), 69–77. https://doi.org/10.1061/(ASCE)1527-6988(2007)8:3(69)

Demuth, J. L., Morss, R. E., Morrow, B. H., & Lazo, J. K. (2012). Creation and communication of hurricane risk information. *Bulletin of the American Meteorological Society*, *93*(8), 1133–1145. https://doi.org/10.1175/BAMS-D-11-00150.1

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A Review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society*, *20*(5), 325–344. https://doi.org/10.1080/01972240490507974

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Garcia-Retamero, R., Cokely, E. T., Ghazal, S., & Joeris, A. (2016). Measuring graph literacy without a test: A brief subjective assessment. *Medical Decision Making*, *36*(7), 854–867. https://doi.org/10.1177/0272989X16655334

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183): North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hohenstein, S., Matuschek, H., & Kliegl, R. (2017). Linked linear mixed models: A joint analysis of fixation locations and fixation durations in natural reading. *Psychonomic Bulletin & Review*, *24*(3), 637–651. https://doi.org/10.3758/s13423-016-1138-y

Huang, S. K., Lindell, M. K., Prater, C. S., Wu, H. C., & Siebeneck, L. K. (2012). Household evacuation decision making in response to Hurricane Ike. *Natural Hazards Review*, *13*(4), 283–296. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000074

IBM Corp (2020). *IBM SPSS statistics for windows, Version 27.0*. IBM Corp.

Lewis, J. R. (1993). Pairs of Latin squares that produce diagram-balanced Greco-Latin designs: A basic program. *Behavior Research Methods, Instruments, & Computers*, *25*(3), 414–415. https://doi.org/10.3758/BF03204534

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099–2102).

Liu, L., Mirzargar, M., Kirby, R. M., Whitaker, R., & House, D. H. (2015). Visualizing time-specific hurricane predictions, with uncertainty, from storm path ensembles. *Computer Graphics Forum*, *34*(3), 371–380. https://doi.org/10.1111/cgf.12649

McNoldy, B. (2022, April 4). "Cone of Uncertainty" Update and Refresher. *Tropical Atlantic Update*. https://bmcnoldy.blogspot.com/2022/04/2022-cone-of-uncertainty-update.html

Millet, B., Carter, A. P., Broad, K., Cairo, A., Evans, S. D., & Majumdar, S. J. (2020). Hurricane risk communication: Visualization and behavioral science concepts. *Weather, Climate, and Society*, *12*(2), 193–211. https://doi.org/10.1175/WCAS-D-19-0011.1

Millet, B., Cairo, A., Majumdar, S., Diaz, C., Evans, S., Broad, K. (2020, October 25–26). Beautiful visualizations slain by ugly facts: Redesigning the National Hurricane Center's 'cone of uncertainty' map. IEEE Visualization for Communication Workshop. https://osf.io/wzk8p/

National Hurricane Center (2018). Florence graphics archive: 5-day forecast track and watch/warning graphic. NOAA/National Weather Service. https://www.nhc.noaa.gov/archive/2018/FLORENCE_graphics.php?product=5day_cone_no_line)

Padilla, L. M., Ruginski, I. T., & Creem-Regehr, S. H. (2017). Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, *2*(1), 40. https://doi.org/10.1186/s41235-017-0076-1

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Ruginski, I. T., Boone, A. P., Padilla, L. M., Liu, L., Heydari, N., Kramer, H. S., Hegarty, M., Thompson, W. B., House, D. H., & Creem-Regehr, S. H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, *16*(2), 154–172. https://doi.org/10.1080/13875868.2015.1137577

Wu, H. C., Lindell, M. K., Prater, C. S., & Samuelson, C. D. (2014). Effects of track and threat information on judgments of hurricane strike probability. *Risk Analysis*, *34*(6), 1025–1039. https://doi.org/10.1111/risa.12128

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the Subjective Numeracy Scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making*, *27*(5), 663–671. https://doi.org/10.1177/0272989X07303824

About the authors

Barbara Millet is a director of the University of Miami User Experience Lab and Assistant Professor of Interactive Media at the School of Communication. Her research focuses on designing products that improve communication of risk and health information in support of human performance in safety-critical settings.

Sharanya J. Majumdar is a professor in Atmospheric Sciences at the University of Miami. His research interests include predictability and processes related to tropical cyclone motion, formation, and intensity change, together with evaluation, data assimilation, observing system design, ensemble prediction, and risk communication.

Alberto Cairo is an associate professor and the Knight Chair in Visual Journalism at the University of Miami. He's the author of several books about data visualization, such as "How Charts Lie" (2019) and "The Truthful Art" (2016). He has a professional background in journalism and visualization design.

Brian D. McNoldy is a senior research associate at the University of Miami's Rosenstiel School of Marine and Atmospheric Science and has been involved in various aspects of hurricane research for over two decades. Research topics have included hurricane dynamics, modeling, observations, climatology, risk perception, and hazard communication.

Scotney D. Evans is an associate professor in the School of Education and Human Development at the University of Miami. He is a community psychologist and community-engaged researcher working to understand and support the role of community-based organizations, networks, and coalitions in building capacity to promote community well-being and social justice.

Kenneth Broad is an environmental anthropologist who studies the use and misuse of scientific information. He is a Professor in the Department of Environmental Science and Policy at the University of Miami where he also serves as Director of UM's Abess Center for Ecosystem Science and Policy.