Spatial Cognition & Computation

An Interdisciplinary Journal

ISSN: 1387-5868 (Print) 1542-7633 (Online) Journal homepage: http://www.tandfonline.com/loi/hscc20

Non-expert interpretations of hurricane forecast uncertainty visualizations

Ian T. Ruginski, Alexander P. Boone, Lace M. Padilla, Le Liu, Nahal Heydari, Heidi S. Kramer, Mary Hegarty, William B. Thompson, Donald H. House & Sarah H. Creem-Regehr

To cite this article: Ian T. Ruginski, Alexander P. Boone, Lace M. Padilla, Le Liu, Nahal Heydari, Heidi S. Kramer, Mary Hegarty, William B. Thompson, Donald H. House & Sarah H. Creem-Regehr (2016) Non-expert interpretations of hurricane forecast uncertainty visualizations, Spatial Cognition & Computation, 16:2, 154-172, DOI: 10.1080/13875868.2015.1137577

To link to this article: http://dx.doi.org/10.1080/13875868.2015.1137577

Accepted author version posted online: 05 Jan 2016.
Published online: 05 Jan 2016.

Submit your article to this journal

Article views: 121

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=hscc20

Download by: [174.52.203.22] **Date:** 14 September 2016, At: 20:09

SPATIAL COGNITION & COMPUTATION
2016, VOL. 16, NO. 2, 154–172
http://dx.doi.org/10.1080/13875868.2015.1137577

Taylor & Francis
Taylor & Francis Group

Non-expert interpretations of hurricane forecast uncertainty visualizations

Ian T. Ruginski,[a] Alexander P. Boone,[b] Lace M. Padilla,[a] Le Liu,[c] Nahal Heydari,[b] Heidi S. Kramer,[d] Mary Hegarty,[b] William B. Thompson,[e] Donald H. House,[c] and Sarah H. Creem-Regehr[a]

[a]Department of Psychology, University of Utah, Salt Lake City, UT, USA; [b]Department of Psychological & Brain Sciences, University of California Santa Barbara, Santa Barbara, CA, USA; [c]School of Computing, Clemson University, Clemson, SC, USA; [d]Division of Epidemiology, University of Utah, Salt Lake City, UT, USA; [e]School of Computing, University of Utah, Salt Lake City, UT, USA

ABSTRACT

Uncertainty represented in visualizations is often ignored or misunderstood by the non-expert user. The National Hurricane Center displays hurricane forecasts using a track forecast cone, depicting the expected track of the storm and the uncertainty in the forecast. Our goal was to test whether different graphical displays of a hurricane forecast containing uncertainty would influence a decision about storm characteristics. Participants viewed one of five different visualization types. Three varied the currently used forecast cone, one presented a track with no uncertainty, and one presented an ensemble of multiple possible hurricane tracks. Results show that individuals make different decisions using uncertainty visualizations with different visual properties, demonstrating that basic visual properties must be considered in visualization design and communication.

KEYWORDS

decision making; hurricane prediction; uncertainty; visualization

1. Introduction

Uncertainty is inherent to predictive models that are used in many real-world applications such as weather forecasts, financial markets, and air traffic control.

However, one significant problem in the *cognition* of uncertainty involves the need to address how uncertainty information can be communicated so the uncertainty can be understood and used most effectively. Our focus in this article is on cognition of the *visual display of uncertainty* in the context of hurricane forecasts. The National Hurricane Center (NHC) displays a track forecast cone, commonly called a "cone of uncertainty" (shown in Figure 1) to communicate a hurricane forecast, including a centerline indicating the predicted hurricane track.

The width of the cone indicates uncertainty in the prediction, based on a five-year sample of historical forecast errors. Specifically, the following is the description of the cone on the NHC website:

CONTACT Ian T. Ruginski Ian.Ruginski@psych.utah.edu Department of Psychology, University of Utah, 380 S. 1530 E., Room 502, Salt Lake City, UT 84112.

© 2016 Taylor Francis

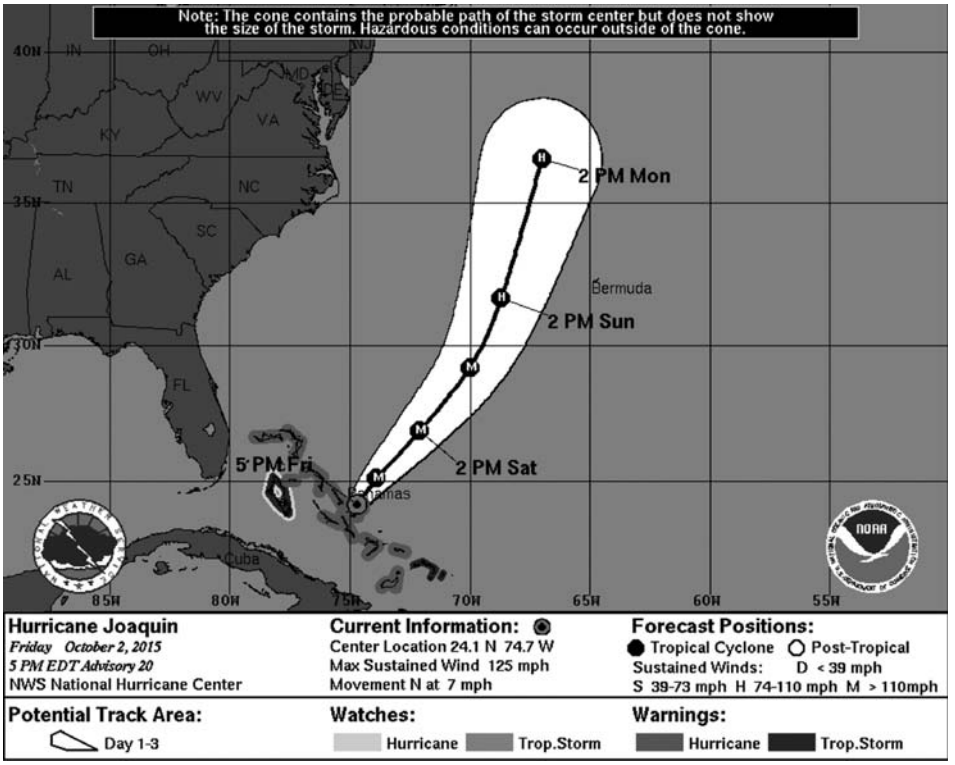

Figure 1. Example of a hurricane forecast cone typically presented to end-users by the National Hurricane Center (http://www.nhc.noaa.gov/aboutcone.shtml).

> The cone represents the probable track of the center of a tropical cyclone, and is formed by enclosing the area swept out by a set of circles (not shown) along the forecast track (at 12, 24, 36 hours, etc). The size of each circle is set so that two-thirds of historical official forecast errors over a 5-year sample fall within the circle." (http://www.nhc.noaa.gov/aboutcone.shtml).

The cone graphically displays information about the probable path over time of the center of the storm, but the shape of the cone says nothing about either the size or intensity of the storm. Because of this, in its standard form, the cone is augmented by annotation of its centerline at regular intervals to indicate the time profile of projected storm intensity. Furthermore, there is no information on the likelihood of any specific path, other than the historical fact that the circles from which the cone is derived encompass two thirds of the actual data from the preceding five years. It is also the case that the historical data predicts that roughly a third of the time the storm center will end up outside the forecast cone.

Studies suggest that this visualization (or simplifications of this visualization used in many media outlets) is associated with misinterpretations that are inconsistent with the data on which the forecast is based, such as a sense of little or no risk for locations outside of the cone (Broad, Leiserowitz, Weinkle, & Steketee 2007; Cox, House, & Lindell, 2013; Wu, Lindell, Prater, & Samuelson,

2014). There is also anecdotal evidence that viewers misinterpret the visual size of the cone as indicating the size or intensity of the storm when in fact it provides no information about size or intensity. Given the apparent saliency of visual information in guiding users' (mis)-understandings of the forecast, we aimed to test whether varying the methods of visualizing uncertainty would influence the intuitive interpretation of hurricane forecasts in a non-expert population.

Communicating uncertainty in visualizations is a broad and generalizable problem. In psychological science, one example that has received much attention is the interpretation of confidence intervals representing uncertainty. Several studies have shown that both students and advanced researchers misunderstand the distribution underlying the confidence interval to be uniform (e.g., Belia, Fidler, Williams, & Cumming, 2005; Zwick, Zapata-Rivera, & Hegarty, 2014). Recent approaches have modified confidence intervals with different graphical encodings such as gradients or violin-like shapes to give more information about the distribution, finding some improvements compared to bar charts with error bars (Correll & Gleicher, 2014), but some work has found little effect of additional visualization of the underlying distribution (Padilla et al., 2015). Other research has focused on types of visual properties that might be more or less intuitively understood as representing uncertainty. For example, MacEachren, Roth, O'Brien, Swingley and Gahegan (2012) found that users rated fuzziness, location, value, arrangement, size and transparency as highly intuitive for depicting uncertainty, but saturation as nonintuitive (see also Howard & MacEachren,1996; Jiang, Ormeling, & Kainz, 1995; Pang, Wittenbrink, & Lodha, 1997). Yet these encodings of uncertainty are still often misinterpreted by end users (Schweizer & Goodchild, 1992; Padilla et al., 2015).

While early work on visualization of geospatial uncertainty focused on developing typologies of uncertainty visualization (e.g,, Buttenfield & Beard, 1991; MacEachren et al., 2005), more recently, a variety of empirical methods, including controlled laboratory studies, have been used to evaluate effectiveness of visualizations of uncertainty (for a recent review see Kinkeldy, MacEachren, & Schiewe, 2014). The tasks in these studies include *subjective judgments*, such as ratings of preference for one visualization over another (e.g., Gerharz & Pebesma, 2009; Senaratne, Gerharz, Pebesma, & Schwering, 2012), or ratings of the intuitiveness of various symbols and *objective* measures, such as time and accuracy to retrieve specific values from a visualization (e.g., Drecki, 2002; Finger & Bisantz, 2002). Interestingly, in their review of 44 studies, Kinkeldy et al. (2014) reported that most of these empirical studies focused on attribute uncertainty, only one focused on positional uncertainty (Grigoryan & Rheingans, 2004) and none focused on temporal uncertainty. The present study follows an increasing trend to use objective measures and controlled laboratory studies to study the cognitive comprehension of uncertainty, and filling a gap in

the empirical literature by examining a case in which the uncertainty is both positional and temporal in nature.

From a cognitive perspective, visualizations such as the hurricane track forecast cone that depict salient size and shape of the cone are likely to be categorized and interpreted for their spatial features, rather than for their statistical properties (Newman & Scholl, 2014). For example, using qualitative methods, Broad et al. (2007) determined that Florida residents and media reporters believed that the NHC track forecast cone represented an area of impact rather than a distribution of potential hurricane tracks and focused on the hurricane centerline as the primary location of hurricane impact. *Ensemble* representations of data, which show many predicted instances, have been suggested as a solution to address the shortcomings of visually presenting discrete spatial properties to depict probabilistic data (Stephens, Edwards, & Demeritt, 2012). Cox et al. (2013) examined whether an ensemble visualization compared to a track forecast cone would better communicate the uncertainty intended by the hurricane forecast, using a probability estimation task. They provided initial evidence that the ensemble visualization may help users better understand the variance inherent in the visualized hurricane forecast.

While the results of Cox et al. (2013) were promising, their study was limited by the use of a probabilistic task and the inclusion of only two types of visualizations. Previous research across a variety of domains and tasks shows that people are poor at probabilistic reasoning and problem-solving (Murphy, Lichtenstein, Fischhoff, & Winkler, 1980; Tversky & Kahneman, 1983). In the current experiment, we sought to expand upon Cox et al.'s (2013) work by creating a task that asked viewers to rate the amount of damage predicted to occur at a given location. This type of objective assessment rating task is increasingly used to evaluate alternative visualizations of geospatial uncertainty (Kinkeldey et al., 2014). Our intent was to use a task that measured the intuitive nature of the uncertainty visualization without presuming knowledge about the meaning of the track forecast cone. Because of our focus on intuitive understanding, we presented our visualizations without an accompanying legend, as is typical in many media outlets, such as TV news.

We predicted that ensemble visualizations would alert users to the presence of uncertainty in the visualization and reduce the misinterpretations of the track forecast cone that could result from the properties of the salient center track and the discrete shape and size of the cone. To further test the role of these cone properties, we created three additional visualizations. A cone with a soft boundary allowed us to examine whether "fuzziness" would convey a sense of uncertainty (e.g., Boukhelifa, Bezerianos, Isenberg, & Fekete, 2012; MacEachren et al., 2012). A cone without a centerline aimed to address whether the absence of a hurricane "track" would change interpretation at or near the center of the cone. Finally, a visualization with only a centerline would test user interpretation of the hurricane forecast in the absence of visually

presented uncertainty. Overall, we predicted that changing the visual properties representing uncertainty would influence cognitive interpretations of the hurricane forecast and that this would be revealed in the level of damage intensity ratings given by users over visually-depicted time and space.

2. Methods

2.1. Participants

Participants were 103 students from the University of Utah and 102 students from the University of California Santa Barbara enrolled in introductory psychology classes. Participants were provided class credit for participating. They were between 17 and 48 years of age with a mean age of 21.3 years. 144 were female, and 60 were male. The gender of one participant was not recorded. Five participants were excluded from final data analyses (one due to experimenter error, one non-native English speaker, one had a color deficiency, and two did not follow directions). A total of 200 participants were included in the final data analysis.

2.2. Stimuli

Stimuli were presented on an Asus PA246 Series LCD monitor and an Asus VG248 monitor in sRGB color mode, using E-Prime 2.0.10.353 software (Schneider, Eschman, & Zuccolotto, 2012). On each trial, participants were presented with a display depicting a hurricane forecast. Five visualization methods were utilized: cone-centerline, centerline-only, ensemble, fuzzy-cone, and cone-only (see Figure 2). The hurricane images were generated using prediction advisory data from historical hurricanes, available on the NHC website (http://www.nhc.noaa.gov/archive). Custom computer code was written to reconstruct uncertainty cones in the required formats, using the same cone construction algorithm described on the NHC website (http://www.nhc.noaa.gov/aboutcone.shtml). The ensemble visualizations were constructed using the ensemble generation code of Cox et al. (2013). All were digitally composited over a map of the U.S. Gulf Coast that had been edited to minimize distracting labeling. These images were displayed to the subjects at a pixel resolution of 1177×1000. A single location of an "oil rig" depicted as a red dot was superimposed on the image at one of twelve locations defined relative to the center track (0) and the cone boundaries (± 1). We chose the following relative points ± 1.9, 1.5, 1.2, 0.8, 0.5, 0.2, at each timepoint with respect to the boundaries of the cone (see Figure 3).

Relative points with respect to the center and cone boundary were chosen so that three points fell outside the cone boundary, (1.9, 1.5, and 1.2) and three points fell within the cone boundary (0.8, 0.5, and 0.2) and so that no points appeared to touch the visible center line or boundary lines. Underneath the

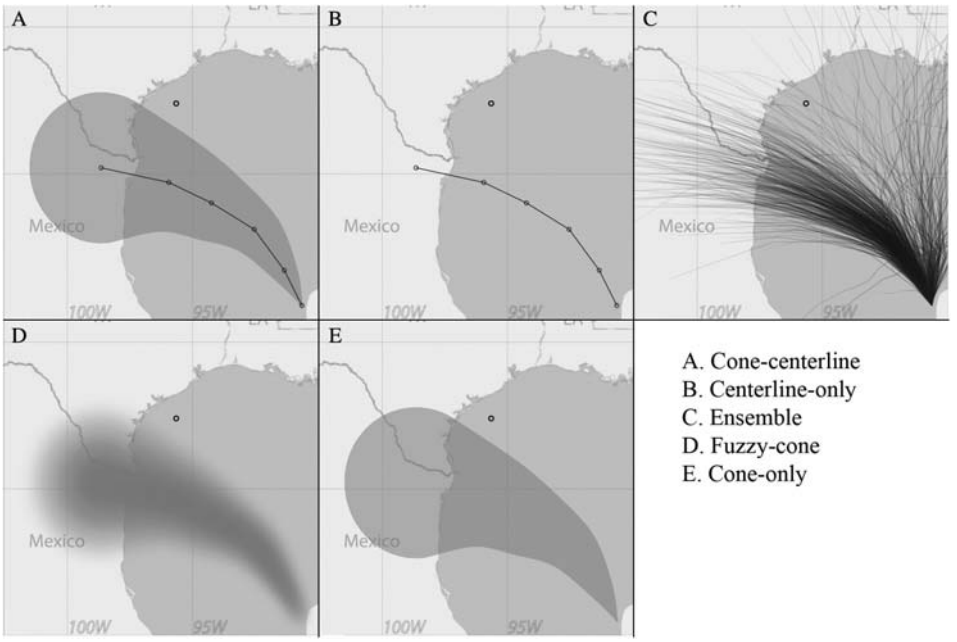

Figure 2. One example of each of the visualizations (presented in color in the actual study with dark blue track lines, light blue cones, and red oil rig locations).

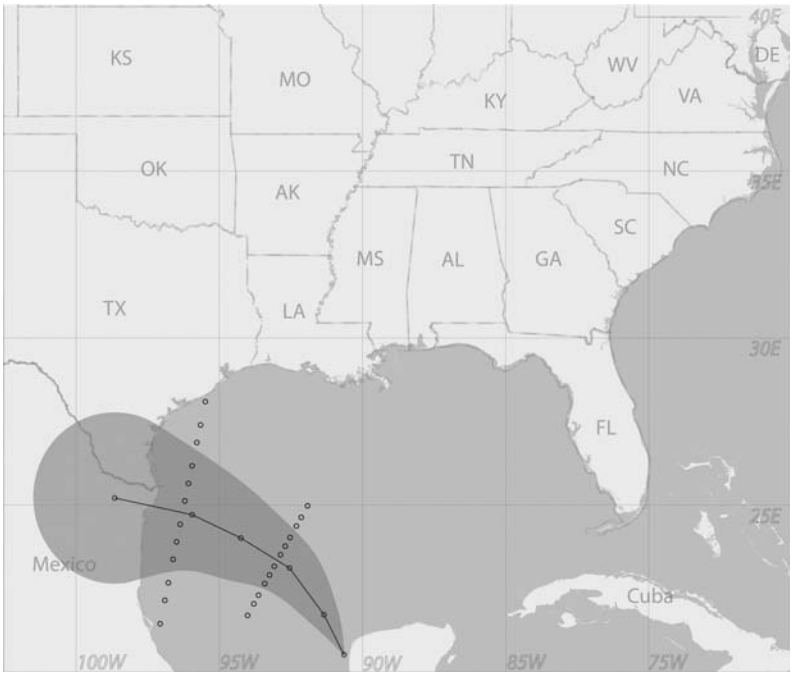

Figure 3. Example of the full display including the cone-centerline visualization, shown with the 12 possible oil rig locations at the 24-hour and 48-hour timepoints. Only one location was presented on each trial.

forecast, a Likert scale ranging from 1 (no damage) to 7 (severe damage) was displayed. This scale was chosen to provide participants with a means of hurricane interpretation without forcing them to directly make probabilistic inferences. Our intent was to assess an initial intuitive understanding of each visualization, given no additional information or instructions. Following the main experiment, an online questionnaire including nine true/false questions assessed students' knowledge of and beliefs concerning hurricanes and the visualizations that they had just seen (see Table 7).

2.3. Design

A 6 (hurricane forecasts) × 12 (oil rig location values) × 2 (24-hour and 48-hour timepoint) × 5 (visualization method) mixed factorial design was used. Participants were randomly assigned to one of five visualization conditions as a between-subjects factor. Hurricane forecast, oil rig location and timepoint were within-subjects variables, resulting in a total of 144 experimental trials per participant.

2.4. Procedure

The lights in the experiment room were turned on. At the University of Utah site, the window blinds were shut to reduce screen glare and at UC Santa Barbara the experiment was conducted in a windowless room. An experimenter ran contrast and gamma monitor tests prior to the task (http://www.lagom.nl/lcd-test/), to ensure display constancy across monitors. Each participant gave informed consent and then was screened for color deficiency. The desk chair and monitor heights were adjusted so that the participant's eye height was at the mid-point of the screen. Each session lasted approximately 1 hour.

The participants were given the following description of the scenario/task and instructions about the uncertainty visualizations presented:

> In the following experiment, you will view maps showing the forecast path of different hurricanes as they travel over the Gulf of Mexico, towards land. The maps will also show the location of one offshore oil platform in the Gulf. Oil platforms are large structures on the surface of the water with components that extend to the ocean floor for drilling and storing oil.

See the sample map below. The forecast path of where the hurricane will move in the next three days is shown in blue and the location of the oil platform is shown by a small red circle. Your task is to estimate the level of damage that the platform will incur based on the depicted forecast of the hurricane path on a scale of 1 to 7 where 1 is no damage and 7 is severe damage.

You will make your judgments of potential damage to the oil platform using the damage scale provided below the map, which will be presented to you along with the forecast maps on each trial. To respond you should press the specific key (1 through 7) associated with the level of damage that you believe will occur

to the oil platform as a result of the forecasted hurricane. The hurricane forecasts and the locations of the oil platforms will vary across trials."

Prior to beginning the task, participants completed three practice trials and indicated to the experimenter if there were any questions. Participants then began the experimental trials. Participants were allowed to take a break once they were halfway done with the task.

Following completion of the task, participants were asked to think aloud while completing 12 additional trials. These trials were chosen to be representative of the overall distribution of oil rig locations and to include points both inside and outside the cone of uncertainty in order to better identify participants' decision-making strategies. Participants were asked to make a damage judgment, and describe the strategies used to inform decision making aloud to the experimenter. The experimenter recorded participants' responses using an Olympus DP-201 voice recorder or a Samsung Galaxy player 4.0. Recorded responses were later transcribed into written text. These self-reports were intended to provide insight into the heuristics informing participants' reasoning process. Following the think-aloud participants completed the online questionnaire.

3. Results

3.1. Analysis of damage judgments

Given the known absolute radii at each timepoint (24-hour = 186.07 km, 48-hour = 347.28 km), we calculated the distances of the oil rig locations from the centerline and used these in the analyses. In addition, we analyzed the absolute value of oil rig distances, regardless of which side of the hurricane forecast they were on, as none of our hypotheses related to whether oil rigs were located on a particular side and because raw data plots revealed slight to no skew in the data distributions (see Figure 4).

A multilevel model was fitted to the data using Hierarchical Linear Modeling (HLM 7.0) software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). Multilevel modeling is a generalized form of linear regression that is used to analyze variance in experimental outcomes on both individual (within-subjects) and group (between-subjects) levels. This model allowed for the inclusion of interactions between continuous variables (in our case, distance) and categorical predictors (in our case, glyph and time point) using estimation procedures appropriate for mixed, nested designs (i.e., repeated measurements nested within an individual; individuals nested within groups). Damage rating, although an ordinal variable by definition, was treated as continuous in the model because it contained over five response categories (Bauer & Sterba, 2011). Assumptions of normality and homoscedasticity of residuals were checked and met.

The mixed two-level regression model tested whether the effect of distance from hurricane center, and the effect of timepoint, as well as their interaction

(level 1 variables), varied as a function of visualization (level 2)[1]. We hypothesized that if participants use the discrete cone shape and size and the center line to guide their interpretation of the intensity of the hurricane, then oil rig distance from centerline would have a greater effect on damage judgments (showing a steep decreasing slope in ratings as a function of distance) for the cone-centerline visualization than for the visualizations that did not include a center line or hard boundaries (cone-only, fuzzy- cone, ensemble). Specifically, given the nature of the ensemble visualization, which provided many instances of hurricane tracks, we expected the smallest effect of distance (i.e., shallowest slope) for the ensemble. In addition, we predicted that ratings of damage would differ as a function of timepoint for the visualizations with a cone versus those without a cone. If the widening of the cone is interpreted as the storm growing over time, then we would expect to see an increase in damage ratings at the 48-hour timepoint with the cone, but no increase with the ensemble or the centerline-only visualization.

We first compared each visualization with the cone-centerline at 24 hours at the location of the center (distance = 0) (see Table 1). We chose the cone-centerline as the standard for comparison because it is most similar to current practice in communicating hurricane forecasts, that is, the National Hurricane Center visualization, shown in Figure 1, which is used as the basis for most media communications of hurricane forecasts[2]. Here, participants' mean damage judgments made when viewing the cone-only and fuzzy-cone (dotted gray lines in Figure 4) were lower than damage judgments made using the

1. Timepoint was dummy coded such that the 24-hour timepoint was coded as 0 and 48-hour timepoint was coded as 1. In addition, visualization type was dummy coded such that the cone-centerline was the reference group. For distance, given the relatively large distance between oil rig locations (average 63.76 km and 118.07 km for the 24-hour and 48-hour timepoints, respectively), we divided distance by 10 prior to analysis so that the coefficient would correspond to a 10 km change (rather than a 1 km change). The analysis collapsed over the six hurricane forecasts.

We used the following model:

Model reliability estimate = .992, $\sigma^2 = 1.06667$, $\tau = 0.87893$

$$\begin{aligned} Damage_{ij} = {} & \gamma_{00} + \gamma_{01}{}^{*} Centerline-only_j + \gamma_{02}{}^{*} Ensemble_j + \gamma_{03}{}^{*} Fuzzy-cone_j + \gamma_{04}{}^{*} Cone-only_j + \gamma_{10}{}^{*} Time_{ij} \\ & + \gamma_{11}{}^{*} Centerline-only_j{}^{*} Time_{ij} + \gamma_{12}{}^{*} Ensemble_j{}^{*} Time_{ij} + \gamma_{13}{}^{*} Fuzzy-cone_j{}^{*} Time_{ij} \\ & + \gamma_{14}{}^{*} Cone-only_j{}^{*} Time_{ij} + \gamma_{20}{}^{*} Distance_{ij} + \gamma_{21}{}^{*} Centerline-only_j{}^{*} Distance_{ij} \\ & + \gamma_{22}{}^{*} Ensemble_j{}^{*} Distance_{ij} + \gamma_{23}{}^{*} Fuzzy-cone{}^{*} Distance_{ij} + \gamma_{24}{}^{*} Cone-only_j{}^{*} Distance_j \\ & + \gamma_{30}{}^{*} TimeXDistance_j + \gamma_{31}{}^{*} Centerline-only_j{}^{*} TimeXDistance_{ij} + \gamma_{32}{}^{*} Ensemble_j{}^{*} TimeXDistance_j \\ & + \gamma_{33}{}^{*} Fuzzy-cone_j{}^{*} TimeXDistance_j + \gamma_{34}{}^{*} Cone-only_j{}^{*} TimeXDistance_{ij} + u_{0j} + r_{ij} \end{aligned}$$

One of the hurricane stimuli used for the study was later found to begin at a slightly different apex for the ensemble visualization. The model was re-run excluding all trials with this hurricane from each condition ($df = 23783$), however results did not differ from the analysis including all trials. Analysis including all trials is reported.

2. Although choosing a different reference group for statistical comparison would change the specific results of the planned contrasts, the overall take-home message of the analyses would not change.

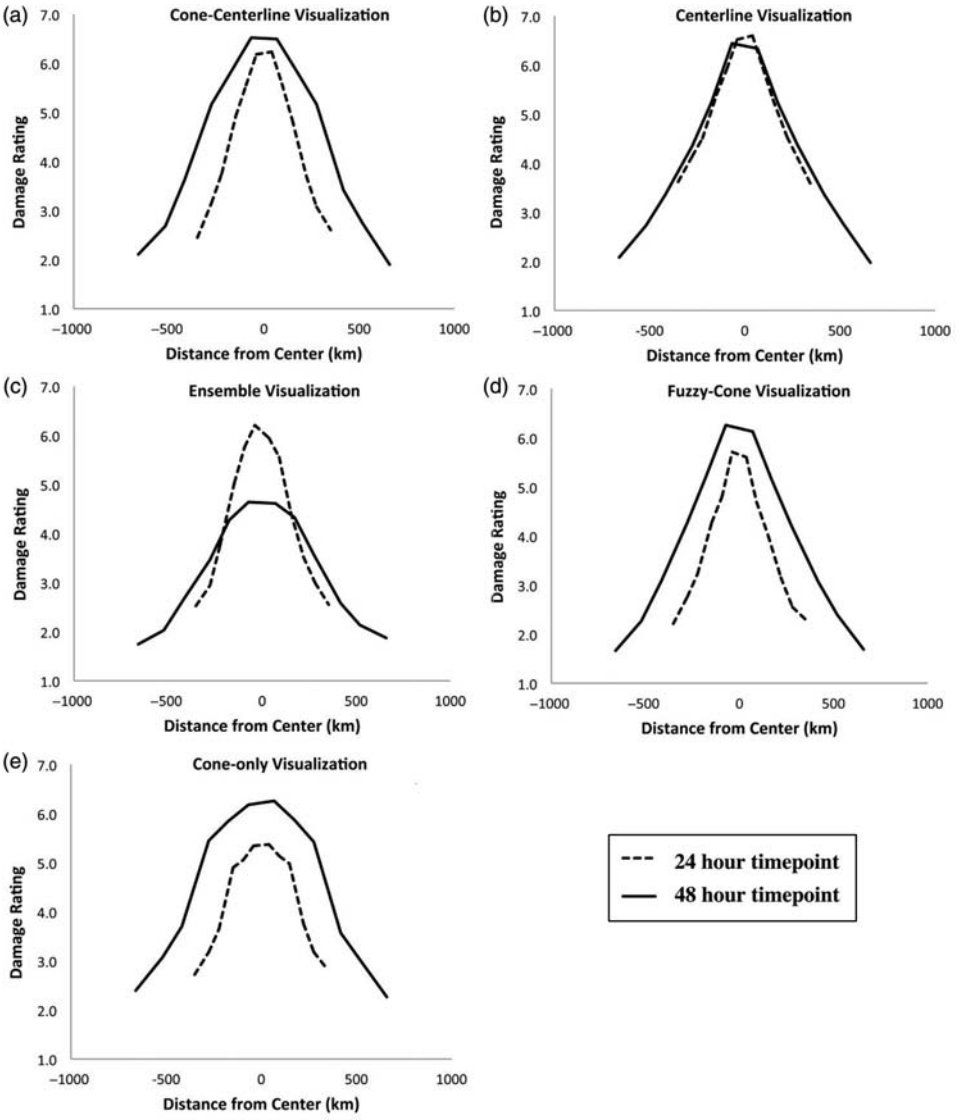

Figure 4. Mean damage ratings as a function of distance from center for each of the 5 visualizations for the 24-hour (dotted lines) and 48-hour (solid lines) timepoints.

Table 1. Damage judgments at the center of the hurricane for the 24-hour timepoint.

Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value	95% CI
For Intercept 1, β_0						
Intercept 2, γ_{00}	6.632	0.15517	42.74	195	<0.001	(6.23, 7.03)
Centerline-only, γ_{01}	0.183	0.22089	0.83	195	0.410	(-0.33, 0.69)
Ensemble, γ_{02}	-0.080	0.21675	-0.37	195	0.711	(-0.59, 0.43)
Fuzzy-cone, γ_{03}	-0.803	0.21675	-3.71	195	<0.001	(-1.40, -0.21)
Cone-only, γ_{04}	-0.744	0.21675	-3.43	195	<0.001	(-1.36, -0.13)

Main effect references cone-centerline group. The table summarizes the comparison of each group to the cone-centerline group.

Table 2. Effect of the change from 24-hour to 48-hour timepoint at the center of the hurricane.

Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value	95% CI
For Time slope, β_1						
Intercept 2, γ_{10}	0.519	0.13877	3.74	28582	<0.001	(0.25, 0.79)
Centerline-only, γ_{11}	-0.748	0.19446	-3.85	28582	<0.001	(-1.13, -0.37)
Ensemble, γ_{12}	-2.071	0.17051	-12.14	28582	<0.001	(-2.41, -1.73)
Fuzzy-cone, γ_{13}	0.179	0.18317	0.98	28582	0.329	(-0.18, 0.54)
Cone-only, γ_{14}	0.561	0.20201	2.78	28582	0.006	(0.16, 0.96)

Main effect references cone with centerline group. The table summarizes the comparison of each group to the cone-centerline group.

cone-centerline (dotted black line in Figure 4). Given that these two visualizations lacked the centerline, this suggests that the centerline may drive participants to believe that the storm is more intense at its middle. Judgments made using the centerline-only and the ensemble visualizations did not significantly differ from judgments made using the cone-centerline at the center of the hurricane for the 24-hour timepoint.

Next, we assessed the difference between the 24-hour and 48-hour timepoints at the center (distance = 0) by looking at the effect of timepoint (see Table 2). Participants in the cone-centerline group made significantly *higher* damage judgments at the 48-hour timepoint (solid black line, Figure 4a) compared with the 24-hour timepoint (dotted black line, Figure 4a) at the center of the hurricane (intercept in Table 2). This finding is consistent with participants using a heuristic that relates size of the cone to the intensity of the storm. Interestingly, the increase in ratings between the timepoints was even greater in the cone-only condition compared with the cone-centerline group, likely due to the lower ratings at 24 hours for the cone-only visualization.

In contrast, participants viewing the ensemble visualization *decreased* in their damage ratings from 24 hours to 48 hours, showing a distinctly different effect of timepoint for the ensemble compared to the cone-centerline (dotted and solid gray lines in the top left of Figure 4). This suggests a different cognitive interpretation of the size or intensity of the hurricane over time when viewing the ensemble. The centerline-only condition also demonstrated a significant but smaller decrease in damage judgments between the 24-hour and 48-hour timepoints, compared to the cone-centerline. Changes in damage judgments between the 24-hour and 48-hour timepoints were similar for the fuzzy-cone and the cone-centerline condition.

Next, we examined the effect of distance from the center at the 24-hour timepoint for each visualization compared to the cone-with-centerline (Table 3). At the 24-hour timepoint, damage judgments in the cone-centerline condition decreased (at a rate of .12 per 10 kilometers) with distance from center of the hurricane (see intercept in Table 3). Both the cone-only and centerline-only conditions slightly reduced this effect, as damage judgments decreased by an average of .09 for the centerline-only and cone-only visualizations. The effect of distance on damage rating did not significantly

Table 3. Effect of a 10 kilometer change in distance at the 24-hour timepoint.

Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value	95% CI
For Distance slope, β_2						
Intercept 2, γ_{20}	−0.121	0.00790	−15.37	28582	<0.001	(−0.13, −0.11)
Centerline-only, γ_{21}	0.026	0.00991	2.63	28582	0.008	(0.01, 0.05)
Ensemble, γ_{22}	0.001	0.00952	0.06	28582	0.950	(−0.02, 0.02)
Fuzzy-cone, γ_{23}	0.013	0.01034	1.22	28582	0.224	(−0.01, 0.03)
Cone-only, γ_{24}	0.031	0.01085	2.83	28582	0.005	(0.01, 0.05)

Main effect references cone with centerline group. The table summarizes the comparison of each group to the cone-centerline group.

Table 4. Interaction between time and distance, i.e, how the effect of distance differs from the 24-hour to 48-hour timepoints.

Fixed effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value	95% CI
For Time × Distance slope, β_3						
Intercept 2, γ_{30}	0.040	0.00489	8.22	28582	<0.001	(0.03, 0.05)
Centerline-only, γ_{31}	−0.018	0.00633	−2.79	28582	0.005	(−0.03, −0.01)
Ensemble, γ_{32}	0.028	0.00612	4.64	28582	<0.001	(0.02, 0.04)
Fuzzy-cone, γ_{33}	−0.009	0.00655	−1.38	28582	0.166	(−0.02, 0.00)
Cone-only, γ_{34}	−0.022	0.00665	−3.27	28582	0.001	(−0.03, −0.01)

Main effect references cone with centerline group. The table summarizes the comparison of each group to the cone-centerline group.

differ between the fuzzy-cone and the cone-with-centerline or between the ensemble and the cone-with-centerline at the 24-hour timepoint. Overall, at this first timepoint, the differences in the effect of distance are relatively small.

However, considering how the effect of distance *changes* from the 24 to 48-hour timepoint shows us important differences among the visualizations (see Table 4). First, the cone-centerline visualization showed less of an effect of distance on damage ratings at the 48-hour timepoint versus the 24-hour timepoint (less negative slope as a function of distance). The effect of distance was reduced even further from 24-hour to 48-hour when viewing the ensemble, compared to the cone-centerline visualization. This suggests that the ensemble leads to an interpretation of damage from the hurricane to be more spatially distributed at the later timepoint compared to the cone-centerline.

Finally, significant variation in damage judgments remained even after controlling for distance from center, timepoint, visualization type, and the interactions, as revealed by the significant random effects, $\chi^2\,(195) = 23330.16$, $p < .001$. This suggests that participants demonstrated significantly different damage judgment means on an individual level, likely as a result of the minimal nature of instructions provided for the task.

3.2. Summary

The analysis of damage ratings provides support for our hypothesis that providing users with an ensemble visualization of multiple instances of hurricane tracks would lead to differences in damage ratings compared to the cone-centerline visualization. Most notably, participants who viewed the

Table 5. Examples of statements that were coded as consistent with each of the heuristics.

Code	Example of statement
Distance	"I would pick a 7 because it's just about, almost really close to the line."
Containment	"A 5 because it's still in the general blue area where the hurricane is heading"
Count	"One wave kind of grazes it, I'm going to go with 1"
Depth of color	"That one is a definite 7 because it's very dark and it's going straight through it"
Curve	"This one I would probably give a 2 even though it's closer to the purple it's on the outside of the curve"
Size	"And I'll give this a 5 because it's towards the bigger part of the storm"
Intensity	"It's like directly in it's path... and it's quite intense, the light there I'm going to go with 7"
Movement	"Seeing as how the hurricane is moving away from the oil rig but there's still a good amount of waves hitting I'm going to go with 2"
Uncodable	"I am going to give this one a 5 because it looks like a 5 to me"

ensemble responded with lower damage ratings as time increased. This was in contrast to the cone, which received higher damage ratings at the later timepoint. This provides further support for the misinterpretation of the cone and directly relates increasing cone size (inherent in the later timepoint) with hurricane intensity/damage.

Furthermore, along with the lower peak damage rating at 48 hours at the center location, there was also less of a change in ratings with increasing distance from the center for the ensemble visualizations. This may be due to the visually dispersed distribution of potential hurricane paths depicted in the ensemble visualizations, as compared to the discrete center line and boundaries depicted in the cone. Finally, there were some smaller differences when comparing the variations of the cone visualizations to the cone-centerline. As depicted in Figure 4, overall higher damage ratings were seen for centerline-only visualization, whereas lower magnitude of ratings were seen for the fuzzy-cone and cone-only, suggesting that varying the visual features communicating uncertainty had some influence, within the context of the cone visualization itself.

3.3. *Think aloud trials*

Participants' 12 think-aloud trials were coded on a trial-by-trial basis to determine which properties or heuristics informed their damage ratings.[3] Participants based their judgments on eight distinct visual properties of the displays (see examples in Table 5). Trials were coded as consistent with a "distance" heuristic when participants mentioned using some distance relationship in the stimuli. "Containment" was coded when participants made judgments based on whether the oil rig was located inside or outside of the blue region. Participants also made judgments based on whether or not the oil rig was on either side of the hurricane track's curve and this was coded as the "curve" heuristic. The code "depth of color" was used when participants

3. Two additional participants were excluded from these analyses due to experimenter error in data collection, so that these and the analyses of questionnaire data were based on 198 participants.

Table 6. Mean number (standard deviation in parentheses) of each heuristic mentioned by visualization condition.

	Ensemble	Cone-only	Fuzzy-one	Cone- centerline	Centerline- only
Distance	4.25 (3.12)	6.83 (2.74)	7.10 (3.45)	7.67 (3.06)	9.76 (2.05)
Containment	2.23 (2.87)	5.00 (3.03)	3.76 (3.23)	4.92 (3.22)	0.55 (1.39)
Count	4.15 (4.38)	0	0	0	0
Depth of Color	1.28 (2.26)	0	1.56 (2.51)	0.03 (0.16)	0
Curve	0	.30 (1.04)	0.32 (0.82)	0.64 (2.13)	1.58 (2.77)
Size	0.02 (0.16)	0.73 (1.65)	0.24 (0.83)	0.79 (1.92)	0.03 (0.16)
Intensity	0.67 (1.40)	0	0.17 (0.67)	0.05 (0.22)	0
Movement	0.53 (1.04)	0.15 (0.48)	0.54 (1.16)	0.36 (1.31)	0.39 (1.10)

made explicit mention of the darkness/lightness of the blue region. When participants spoke about counting the number of lines that touched the oil rig in the Ensemble condition, we assigned the "count" code. We also coded for mention of the movement of the hurricane and for judgments based on the size and intensity of the hurricane, properties that were not shown in the visualization. Finally, some statements were uncodable. Interrater reliability (Cohen's Kappa) between two independent coders was 0.84.

Table 6 presents mean numbers of trials on which participants in each condition justified their response with respect to each coded visual property. We conducted nonparametric analyses because these data were not normally distributed. We first compared frequency of mention of the coded visual properties across the five visualization conditions using the Kruskal–Wallis test. This analysis revealed significant differences between conditions for all visual properties except movement ($p < .01$ in all cases).

Next, we compared verbal reports from participants who viewed each visualization condition to those who were in the cone-centerline condition (as in the analyses of the damage rating data). There were no significant differences between cone-only and cone-centerline conditions ($p > .15$ in all cases). In the comparison of the fuzzy-cone and cone-centerline conditions, only depth of color showed a significant difference ($p < .001$) such that the fuzzy-cone participants indicated using this visual property more in their decisions. Significant differences were found between the centerline-only and cone-centerline conditions. Distance and curve were mentioned more by those in the centerline-only condition while containment and size were mentioned more by those in the cone-centerline condition ($p < .03$ in all cases).

Critically, a comparison of the cone-centerline and ensemble visualizations revealed significant differences ($p < .04$ in all cases) for all coded visual properties except movement. Participants in cone-centerline condition reported the distance and containment heuristics more often; the Count heuristic was used exclusively by the ensemble group. The cone-centerline group also had more mentions of the curve and size of the hurricane, whereas the ensemble group more often mentioned intensity and depth of color. In sum, it is clear that

Table 7. Count and percentage of participants who agreed with each statement by condition.

Question	Ensemble	Cone- only	Fuzzy cone	Cone- Centerline	Centerline only
1. The display shows a distribution of possible hurricane tracks	39 (93%)	37 (88%)	38 (90%)	33 (87%)	33 (83%)
2. The blue region shows the area that is likely to be damaged	40 (95%)	40 (95%)	34 (85%)	32 (82%)	25 (63%)
3. The center of the visualization shows where the hurricane is more likely to travel in the next few days.	33 (79%)	35 (83%)	34 (81%)	36 (93%)	33 (83%)
4. Areas not shown in blue are not predicted to be damaged by the hurricane	30 (71%)	16 (38%)	27 (64%)	19 (49%)	11 (28%)
5. The display shows the hurricane getting large over time	13 (31%)	34 (81%)	29 (69%)	27 (69%)	2 (5%)
6. The display shows the extent of the damage of the hurricane getting greater over time	13 (31%)	21 (50%)	18 (43%)	23 (47%)	4 (10%)
7. The visualization shows where the eye of the hurricane is likely to travel over the next three days	30 (71%)	29 (69%)	26 (62%)	33 (85%)	33 (83%)
8. The hurricane is not likely to travel outside the region shown in blue	26 (62%)	20 (48%)	21 (50%)	19 (49%)	16 (40%)
9. The display indicates that the forecasters are less certain about the path of the hurricane as time passes	23 (55%)	11 (26%)	18 (43%)	20 (51%)	10 (25%)

the various visualizations caused participants to notice different visual properties of the displays and to base their judgments on different heuristics

3.4. Posttask questionnaire

Table 7 provides each statement in the post-task questionnaire, along with the count and percentage of participants who agreed with each statement. Chi-square tests of independence were conducted to evaluate the relationship between visualization condition and endorsement of each of the statements. Again, we conducted comparisons of the cone-centerline condition with the other four conditions.

Compared to the cone-centerline condition, participants in the cone-only condition were less likely to agree that the forecasters are less certain of the forecast as time passes (Statement 9), χ^2 (1, $N = 79$) = 5.79, $p = .02$, and participants in the fuzzy-cone condition were less likely to agree that the visualization shows where the eye of the hurricane is likely to travel (Statement 7), χ^2 (1, $N = 80$) = 4.64, $p = .03$. There were no other differences between the cone conditions.

Not surprisingly, compared to the cone-centerline condition, participants in the centerline-only conditions were less likely to believe that the blue region shows the area that is likely to be damaged (Statement 2), χ^2 (1, $N = 77$) = 4.37, $p = .04$. They were also less likely to agree that the hurricane gets larger over time (Statement 5), χ^2 (1, $N = 77$) = 33.54, $p < .001$ that the extent of the damage becomes greater over time (Statement 6), χ^2 (1,

$N = 77) = 22.54$, $p < .001$, and that forecasters are less certain about the path of the hurricane as time passes (Statement 9), χ^2 (1, $N = 77) = 5.05$ $p = .03$.

Importantly, compared to the cone-centerline condition, participants in the ensemble condition were less likely to endorse the statements that the hurricane would get larger over time (Statement 5), χ^2 (1, $N = 79) = 10.66$, $p < .001$ and that the damage would increase over time (Statement 6), χ^2 (1, $N = 79) = 6.72$, $p = .01$. These results are consistent with the quantitative damage judgments.

Taken together, the think aloud and questionnaire results indicate that the ensemble hurricane visualization reduced the reliance on distance to center and size of hurricane heuristics, consistent with the analyses of damage ratings.

4. Discussion

Broadly, we examined whether varying the visual properties of a hurricane forecast with uncertainty would change users' interpretation of the forecast. While previous work within the domain of hurricane forecasts suggested that non-experts misinterpret the standard track forecast cone presented by the NHC, much of this evidence has been qualitative or anecdotal. Our goal was to build on recent efforts to examine a conceptually different type of visualization, ensemble visualizations (Cox et al., 2013), which may be less susceptible to cognitive misinterpretation because they more explicitly provide multiple instances of predicted hurricane tracks. Our approach was to create five different visualizations—including an ensemble as well as four variations of the presence and nature of the track forecast cone. We developed a "layperson" task to assess the amount of damage that would result at specified locations with respect to the visualized forecast. We also collected extensive qualitative data while users viewed specific visualization trials and more generally on users' beliefs about what the visualization represented.

Our results support the prediction that visual properties matter when considering the cognition of uncertainty visualization. Although the larger radius of the NHC cone (with increasing time into the future) is intended to communicate an increase in uncertainty in the track prediction, it appears that it is cognitively difficult to resist associating the perceived size with the actual size or intensity of the hurricane. When people were presented with the ensemble, which lacks the salient increase in cone size with time, the misconception that the hurricane grows larger or more intense over time was reduced. This was revealed in the decreased damage ratings at 48 hours, and through the significantly lower percentage of responses supporting statement 5 ("hurricane getting larger over time") and statement 6 ("extent of the damage of the hurricane getting greater over time) for the ensemble compared to the cone-centerline.

Furthermore, while distance to the center (defined by the location at the center track) is clearly a valuable heuristic and we found evidence for its use in both the damage ratings and think aloud judgments, we also found a difference

in use of this heuristic in the ensemble compared to the cone-centerline conditions. Notably, at 48 hours, data from the ensemble group revealed a more distributed assessment of damage compared to the sharp decline of damage ratings with distance seen in the other visualizations.

Comparisons of the nonensemble visualizations with the more standard cone-centerline also revealed some interesting differences. The fuzzy-cone and the cone-only visualizations (both without the centerline) resulted in lower damage judgments than the cone-centerline. It is possible that the presence of the salient center forecast track leads users to cognitively assess the intensity of the hurricane to be greater. There was not, however, a strong indication of an effect of the fuzzy boundaries on damage ratings, as the linear effect of distance did not differ between the fuzzy-cone and the cone-centerline. However, descriptively, the pattern of ratings for the fuzzy cone appears to be less influenced by the boundary of the cone than that seen in the cone-only condition. Future work could examine the potential effects of fuzzy versus hard boundaries further. Participants also reported use of different heuristics and visual properties to aid their judgments depending on visualization type.

There are some limitations to the current study that lead directly to future directions of research. First, as mentioned previously, questionnaire data suggested that the ensemble visualization helped reduce the misconception (evident in the cone conditions) that the hurricanes get larger over time. Although the quantitative damage ratings may be related to participants' interpretation of the size of the hurricane, we did not directly measure judged size. To address this question more directly, we have developed a methodology to assess judgments of size and intensity with similar hurricane forecast visualization stimuli. Second, because the goal of this experiment was to study intuitive uncertainty visualization judgments, participants were non-experts who were provided with little information concerning the conventions of the display, as well as no social information (see Meyer, Broad, Orlove, & Petrovic, 2013).

In a follow-up project currently underway, we are examining effects of additional information about display conventions (such as explicit verbal instructions about the meaning of the cone or including a legend) on participants' damage judgments and understanding. On a similar note, the participants in this experiment were novice users who have not had direct experience with hurricanes. Generalizing these results to individuals with direct experience of hurricanes, and to domain experts, will be other useful future directions.

In sum, this study serves as an example of the value of considering cognition in visually-based uncertainty display design. Although the results of this study were limited to hurricane forecast visualizations, we believe that similar approaches and methods could be used to generalize to other geospatial domains involving uncertainty quantification and visualization. Overall, improving interpretations of uncertainty visualizations through cognitively informed display design can

help better inform users' decision making and has the potential to significantly reduce financial and health costs to the general public.

References

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*(4), 373–390.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396.

Boukhelifa, N., Bezerianos, A., Isenberg, T., & Fekete, J. D. (2012). Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2769–2778.

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), 651–667.

Buttenfield, B., & Beard, M. K. (1991). Visualizing the quality of spatial information. *Proceedings of Auto Carto*, *10*(6), 423–427.

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151.

Cox, J., House, D., & Lindell, M. (2013). Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification*, *3*(2), 143–156.

Drecki, I. (2002). Visualization of uncertainty in geographical data. In W. Shi, P. F. Fisher, & M. F. Goodchild (Eds.), *Spatial data quality* (pp. 140–159). London/New York: Taylor & Francis.

Finger, R., & Bisantz, A. M. (2002). Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical Issues in Ergonomics Science*, *3*, 1–25.

Gerharz, L. E., & Pebesma, E. J. (2009). Usability of interactive and non-interactive visualisation of uncertain geospatial information. In *Proceedings of the Geoinformatik* (pp. 223–230), April 2009, Osnabrück, Germany.

Grigoryan, G., & Rheingans, P. (2004). Point-based probabilistic surfaces to show surface Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, *10*, 564–573.

Howard, D., & MacEachren, A. M. (1996). Interface design for geographic visualization: Tools for representing reliability. *Cartography and Geographic Information Science*, *23*(2), 59–77.

Jiang, B., Ormeling, F., & Kainz, W. (1995). Visualization support for fuzzy spatial analysis. In *Proceedings of the American Congress on Surveying and Mapping: American Society for Photogrammetry and Remote Sensing (ACSM/ASPRS Conference for short)* (pp. 291–300), March 1995, Charlotte, North Carolina.

Kinkeldey, C., MacEachren, A. M., & Schiewe, J. (2014). How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, *51*(4), 372–386.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, *32*(3), 139–160.

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2496–2505.

Meyer, R., Broad, K., Orlove, B., & Petrovic, N. (2013). Dynamic simulation as an approach to understanding hurricane risk response: Insights from the Stormview Lab. *Risk Analysis*, *33*(8), 1532–1552.

Murphy, A. H., Lichtenstein, S., Fischhoff, B., & Winkler, R. L. (1980). Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society*, *61*(7), 695–701.

National Weather Service: National Hurricane Center. (2014). *Definition of the NHC Track Forecast Cone*. Retrieved from http://www.nhc.noaa.gov/aboutcone.shtml

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*(4), 601–607.

Padilla, L. P., Hansen, G., Ruginski, I. T., Kramer, H., Thompson, W. B., & Creem-Regehr, S. H. (2015). The influence of different graphical displays on nonexpert decision making under uncertainty. *Journal of Experimental Psychology: Applied*, *21*(1), 37–46.

Pang, A. T., Wittenbrink, C. M., & Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer*, *13*(8), 370–390.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *Hierarchical Linear Modeling (HLM) 7*. Lincolnwood, IL: Scientific Software International Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh, PA: Psychology Software Tools, Inc.

Schweizer, D. M., & Goodchild, M. F. (1992). Data quality and choropleth maps: An experiment with the use of color. *GIS LIS-International Conference*, *2*(American Society for Photogrammetry and Remote Sensing), 686–686, November 1992, San Jose, California.

Senaratne, H., Gerharz, L., Pebesma, E., & Schwering, A. (2012). Usability of spatio-temporal uncertainty visualisation methods. In J. Gensel, D. Josselin, & D. Vandenbroucke (Eds.), *Bridging the Geographic Information Sciences* (pp. 3–23). Berlin/Heidelberg: Springer.

Stephens, E. M., Edwards, T. L., & Demeritt, D. (2012). Communicating probabilistic information from climate model ensembles—Lessons from numerical weather prediction. *Wiley Interdisciplinary Reviews: Climate Change*, *3*(5), 409–426.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

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.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116–138.