Journal of Experimental Psychology: Applied

1076-898X/14/$12.00 http://dx.doi.org/10.1037/xap0000037

The Influence of Different Graphical Displays on Nonexpert Decision Making Under Uncertainty

Lace M. Padilla
University of Utah

Grace Hansen
National Institutes of Health, Bethesda, Maryland

Ian T. Ruginski, Heidi S. Kramer, William B. Thompson, and Sarah H. Creem-Regehr
University of Utah

Understanding how people interpret and use visually presented uncertainty data is an important yet seldom studied aspect of data visualization applications. Current approaches in visualization often display uncertainty as an additional data attribute without a well-defined context. Our goal was to test whether different graphical displays (glyphs) would influence a decision about which of 2 weather forecasts was a more accurate predictor of an uncertain temperature forecast value. We used a statistical inference task based on fictional univariate normal distributions, each characterized by a mean and standard deviation. Participants viewed 1 of 5 different glyph types representing 2 weather forecast distributions. Three of these used variations in spatial encoding to communicate the distributions and the other 2 used nonspatial encoding (brightness or color). Four distribution pairs were created with different relative standard deviations (uncertainty of the forecasts). We found that there was a difference in how decisions were made with spatial versus nonspatial glyphs, but no difference among the spatial glyphs themselves. Furthermore, the effect of different glyph types changed as a function of the variability of the distributions. The results are discussed in the context of how visualizations might improve decision making under uncertainty.

Keywords: uncertainty, decision making, visualization

Though it is inherent to most data, uncertainty information has been classically difficult to capture and display, whether the result of measurement, forecasting, or modeling (Pang, Wittenbrink, & Lodha, 1997). Computational models are now being used to help make far-reaching policy decisions that impact the health, safety, and well-being of many millions of people. These decisions pertain to a wide range of issues including quality of our air and water, the treatment of disease, and the availability and use of energy, as well as transportation, public safety, and economics. Errors and uncertainties are intrinsic to this process and must be accounted for if decisions are to be made in a rational and optimal manner. Despite the importance of conveying uncertainty, graphical displays do not always present uncertainty-based information in a way that is easily interpretable to the intended user, and decision makers do not always utilize uncertainty-based information as it is intended. The visual and spatial design principles influencing uncertainty-based decision making should be carefully considered to generate effective solutions to this problem (Zwick, Zapata-Rivera, & Hegarty, 2014). Our goal was to determine how variations in several relatively simple visual displays presenting varying amounts of uncertainty would inform decisions by nonexperts in a weather forecast scenario.

Current approaches to uncertainty visualization vary in many ways, including the way uncertainty is represented quantitatively. One approach is to represent a range of variability in the visualized data. Existing uncertainty visualization has encoded this type of uncertainty through "graying out" using color saturation (Hengl, 2003), "out of focus" using blur (Jiang, Ormeling, & Kainz, 1995), "foggy" using transparency (Rhodes, Laramee, Bergeron, & Sparr, 2003; MacEachren et al., 2005), "noisy spots" using texture (Howard, & MacEachren, 1996), and "glyphs" which are compound point symbols that encode both data point and uncertainty (Wittenbrink, Pang, & Lodha, 1996; Pang et al., 1997) or glyph "ensembles", which integrate multiple dimensions of data simulations (Sanyal et al., 2010). A second approach is to encode probabilistic forms of uncertainty using displays such as "icon arrays" or visual grids that show the proportion of data falling in certain conditions (Garcia-Retamero & Cokely, 2013).

Others, such as MacEachren et al. (2012), have defined uncertainty into subcomponents (such as spatial and temporal accuracy, precision, and trust), and asked nonexperts to judge intuitiveness

Lace M. Padilla, Department of Psychology, University of Utah; Grace Hansen, Section on Integrative Neuroimaging, Clinical Brain Disorders Branch, National Institutes of Health, Bethesda, Maryland; Ian T. Ruginski and Heidi S. Kramer, Department of Psychology, University of Utah; William B. Thompson, School of Computing, University of Utah; Sarah H. Creem-Regehr, Department of Psychology, University of Utah.

This research is based upon work supported by the National Science Foundation under Grant 1212806. We thank Mary Hegarty and Robert M. Kirby for comments on the manuscript.

Correspondence concerning this article should be addressed to Sarah H. Creem-Regehr, 380 S. 1530 East, Room 502, Department of Psychology, University of Utah, Salt Lake City, UT 84112. E-mail: sarah.creem@psych.utah.edu

of (and make accuracy judgments using) different visual encodings of these subproperties of uncertainty. The logic used is that features that are spontaneously interpreted as conveying uncertainty could be more effectively used in displays. These researchers found that fuzziness, location, value, arrangement, size, and transparency were all rated as highly intuitive, with saturation rated as very low intuitively—conflicting with Hengl (2003) and others' claims above. It appears that there is a lack of a consensus in the visualization literature concerning exactly what visual properties are best for facilitating understanding in nonexpert users.

The types of encodings used and studied in the visualization communities routinely focus on visual metaphors for uncertainty and often neglect the underlying cognitive mechanisms that inform decision making and action. Some work has attempted to empirically evaluate these encodings by simultaneously accounting for visual information as well as the cognitive mechanisms involved in evaluating uncertainty visualizations. In other words, how are users actually interpreting and using these encodings of uncertainty information? Schweizer and Goodchild (1992) tested whether value or saturation better coded uncertainty information on a choropleth (color-coded) map of the United States, finding that both experienced and nonexpert users misunderstood uncertainty information more often when provided with an accurate legend compared with an inverted legend. This provided insight into what Schweizer and Goodchild call a "dark is more" heuristic—darker colors are associated with more certainty—that individuals typically bring to the task based on past experience or domain-specific common knowledge. Aerts, Clarke, and Keuper (2003) also found that encoding higher levels of uncertainty as lighter red and lower levels as darker red was an effective method of communicating uncertainty in an urban planning context. Hengl (2003) proposed a model for encoding continuous uncertainty by using paleness or whiteness. Evans (1997) found that the "dark is more" heuristic also applies to animated visualizations. It appears, then, that these almost predetermined heuristics may have a greater impact on users' interpretations of visualizations of uncertainty than legends or even the visual encodings themselves. More recently, Zuk and Carpendale (2007) also argued that consideration of users' reasoning processes would help generate more easily interpretable uncertainty visualizations. Users themselves, as well as the nature and domain of the task at hand, must be considered in addition to the data. Sanyal, Zhang, Bhattacharya, Amburn, and Moorhead (2009) set out to test some of these principles using different methods of visualizing uncertainty with 1D and 2D data sets, finding that efficiency of interpretation of visual encodings significantly interacts with the nature of the task. Together, these studies suggest that encodings of uncertainty should account for the heuristics that bias both users' reasoning processes and final decisions.

While uncertainty information is important and inherent to data presentation, past research demonstrates that nonexpert decision makers are especially inept at making judgments using visualizations of uncertainty because of misunderstanding of information provided (Joslyn, Savelli, & Nadav-Greenberg, 2011) or cognitive biasing of information (Klayman & Ha, 1987). Oftentimes, these factors contribute to erosion of trust in the data source, which further biases users' decisions that are made using uncertain information (MacEachren et al., 2012). This phenomenon is not unique to weather or map-based applications and has also been noted across other tasks, such as predicted change in financial information over time (Bisantz, Marsiglio, & Munch, 2005), judgments of risk in an economic game (Morone & Ozdemir, 2012), and judgments of air traffic control flow (Masalonis & Parasuraman, 2003). While these issues arise within multiple tasks and domains, we sought to better understand users' decision-making processes related to the uncertainty visualizations inherent to weather forecasting.

Though it appears that uncertainty visualizations matter to decision makers, it is unclear why certain types of visualizations are commonly misinterpreted. In psychology and statistics, for example, bar graphs have been shown to be misinterpreted, with users interpreting data points visually "within" the bar as more likely a part of the underlying distribution than points falling "outside" the bar that were visually equidistant from the mean (Newman & Scholl, 2012). There has been a push to include confidence intervals in order to better guide users' understanding of uncertainty information in these contexts (Cumming & Finch, 2005). Fidler and Loftus (2009) make the argument that confidence intervals should replace *p* values as the accepted form of data presentation in the social sciences due to their ability to better visually represent uncertainty information (related to the underlying distribution of data). Yet empirical work shows that even expert users with experience in statistics and/or psychology perform poorly at statistical inference tasks involving uncertainty-based confidence intervals (Belia, Fidler, Williams, & Cumming, 2005). Because uncertainty visualizations are both commonly misunderstood and often implemented in weather forecasting applications, we sought to test three different visualization glyphs that presented forecast uncertainty spatially.

Research is not consistent regarding what matters when attempting to accurately visually portray domain-specific uncertainty information. The techniques discussed previously that are typically used to portray uncertainty information in the statistics and visualization literatures have not been sufficiently tested in domain-specific applications. Recent work shows that nonexpert users better utilize and trust uncertainty-inclusive predictive interval visualizations of temperature over deterministic (single-value) visualizations of the same underlying temperature data (Joslyn, Nemec, & Savelli, 2013). Despite the issues inherent to uncertainty information display, uncertainty information is still very important to domain-specific applications for conveying information accurately to users and for gaining user trust in order to facilitate safe, rational decision making in the face of extreme weather forecasts. In one sample of real users, Broad, Leiserowitz, Weinkle, and Steketee (2007) showed that hurricane error cones typically used by the National Hurricane Center were frequently misinterpreted, with users believing that the cone represented area of impact rather than potential hurricane tracks. Cox, House and Lindell (2013) proposed to solve this problem by generating an ensemble spaghetti plot that differentially presents uncertainty information to nonexpert users. A better understanding of the cognitive and perceptual mechanisms of decisions related to uncertainty visualization could help lessen costly user misinterpretations in future applications.

The current experiment was designed to examine how changing the visual representation of uncertainty could influence understanding of the information presented, to explore how users' heuristics may bias reasoning processes and final decisions, and finally, to increase understanding of the cognitive and perceptual mechanisms of decisions related to uncertainty visualization. We examined whether different methods of visually depicting the underlying distributions of weather forecasts would influence nonexpert decision making. In this experiment, we asked viewers to perform a classification task—which of two weather forecasts for a specific day was the most accurate prediction of a given temperature value? We represented uncertainty in the forecast as a normal distribution of likely values. By including uncertainty with domain-specific data, we ensured that our decision task required cognition of uncertainty.

We created five different glyphs. Three of the glyphs represented temperature distributions spatially. One was an interval denoting a range of values within which there is a .95 probability that the actual value lies, analogous to a 95% confidence interval for sample means. Two others were variations on the 95% interval intended to provide more information about the underlying distribution (e.g., that values increasingly further from the center had lower probabilities and that there were not discrete boundaries at the end of the interval). The final two glyphs used nonspatial features—brightness and color—to encode the variability in the distribution. We predicted that for a sample of nonexpert decision makers, glyphs portraying more explicit encoding of the underlying distributions might be more likely to lead participants to consider the uncertainty information. If participants do use the uncertainty information, one possible outcome is that they would evaluate a forecast as more accurate based on an interpretation of the actual value as being more likely, where likelihoods are defined by the values of the two probability density functions at the presented temperature value. A second possibility is that participants would make their decisions based on an interpretation of uncertainty as variability in the forecast, rather than likelihood. Variability is characterized using the standard deviation of the distribution. Alternatively, it is possible that participants would ignore uncertainty information altogether and interpret the forecast that is closer to the actual value as being more accurate. In addition, we predicted that the effect of glyph type on the use of uncertainty information would be influenced by the amount of variability in the distribution sets presented.

Method

Participants

Participants were 106 students from the University of Utah participating to fulfill course requirements (age: 18–51, mean age: 22, gender: 67 female, 36 male, 3 unrecorded). All participants were screened for color deficiency. One additional participant was excluded for a color deficiency.

Stimuli

Stimuli were presented on a 558.4 × 380.8 mm Asus PA246 Series LCD Monitor with 1,920 × 1,200 pixel resolution, using E-Prime v2.0 software. On each trial, participants were presented with a display depicting two weather forecasts. These were fictional forecasting systems labeled “StormWatch” and “Weathernet,” which were displayed next to one another horizontally on the screen (see Figure 1). Each forecast used one glyph, which encoded the predicted temperatures in Fahrenheit. Between the two forecasts, a black plus sign was presented that indicated the actual temperature recorded, labeled “Recorded Temperature.”

Stimuli were generated by adopting a Bayesian view of the uncertainty in the forecast data (Marzouk, Najm, & Rahn, 2007). This approach uses probability distributions to model both random events and partial knowledge of the world. In our case, we presumed that uncertainty in the forecast temperature was characterized by a normally distributed likelihood function. We created four pairs of forecasts, each with a difference of means of 2.62° Fahrenheit and varying standard deviations (*SD*) such that one distribution of the pair had a relatively smaller *SD* than the other distribution, while maintaining the same difference between means (see Figure 2 and Table 1). We refer to these as *distribution sets.* For the four distribution sets, the “Recorded Temperature” values shown on individual trials were one of +/−0.26, +/−0.78, +/−1.57, or +/−4.18 °F relative to the average of the two distribution means.

Two additional manipulations were added to create more trials, using the same four distribution sets. The first was to randomly shift the two likelihood distributions and the corresponding “Recorded Temperature” values such that the average of the two distribution means was uniformly distributed over the range 72.23° – 82.68° Fahrenheit. The second was that the two different *SD*s and the two different means were presented on both left and right sides of the display. This gave us four distinct trials for each “Recorded Temperature” value (also referred to as the probe point) associated with each distribution set.

Five glyph types were generated (see Figure 3). The 95% Interval (I95) glyph presented the distribution with a dot representing the mean and a line with end caps encoding a range of values such that there is a .95 probability that the actual value lies within the interval. Prior research suggests that intervals such as this are often misinterpreted as signaling a uniform distribution with well-defined boundaries (Cumming & Finch, 2005; Cumming, 2007). A second glyph (IMulti) attempted to correct this misinterpretation by using no end caps or explicit indication of the mean and stepped intervals representing .47, .95, and .98 probabilities. The Probability Density Function (PDF) glyph explicitly encoded the probability density function, with darker values indicating greater likelihood and lighter values indicating less likelihood. This glyph was intended to visually encourage treatment of the distribution as a nonuniform distribution. The Dot glyph encoded the *SD* by brightness, with darker values corresponding to smaller *SD*. This type of graphic would be appropriate for visualization applications where spatial encoding of distribution values is not feasible. The Colored Dot (CDot) glyph was similar to Dot but encoded the *SD* using a black-body radiation color map (Borland & Taylor, 2007) that also varied in brightness in the same manner as the Dot glyph, providing two separate channels of information. Both Dot and CDot varied continuously as a function of the underlying standard deviation, though only seven examples were given in the legend.

Design

A 4 (distribution set) × 8 (probed temperature values) × 5 (glyph type) mixed factorial design was used. Participants were randomly assigned to one of five glyph type conditions as a

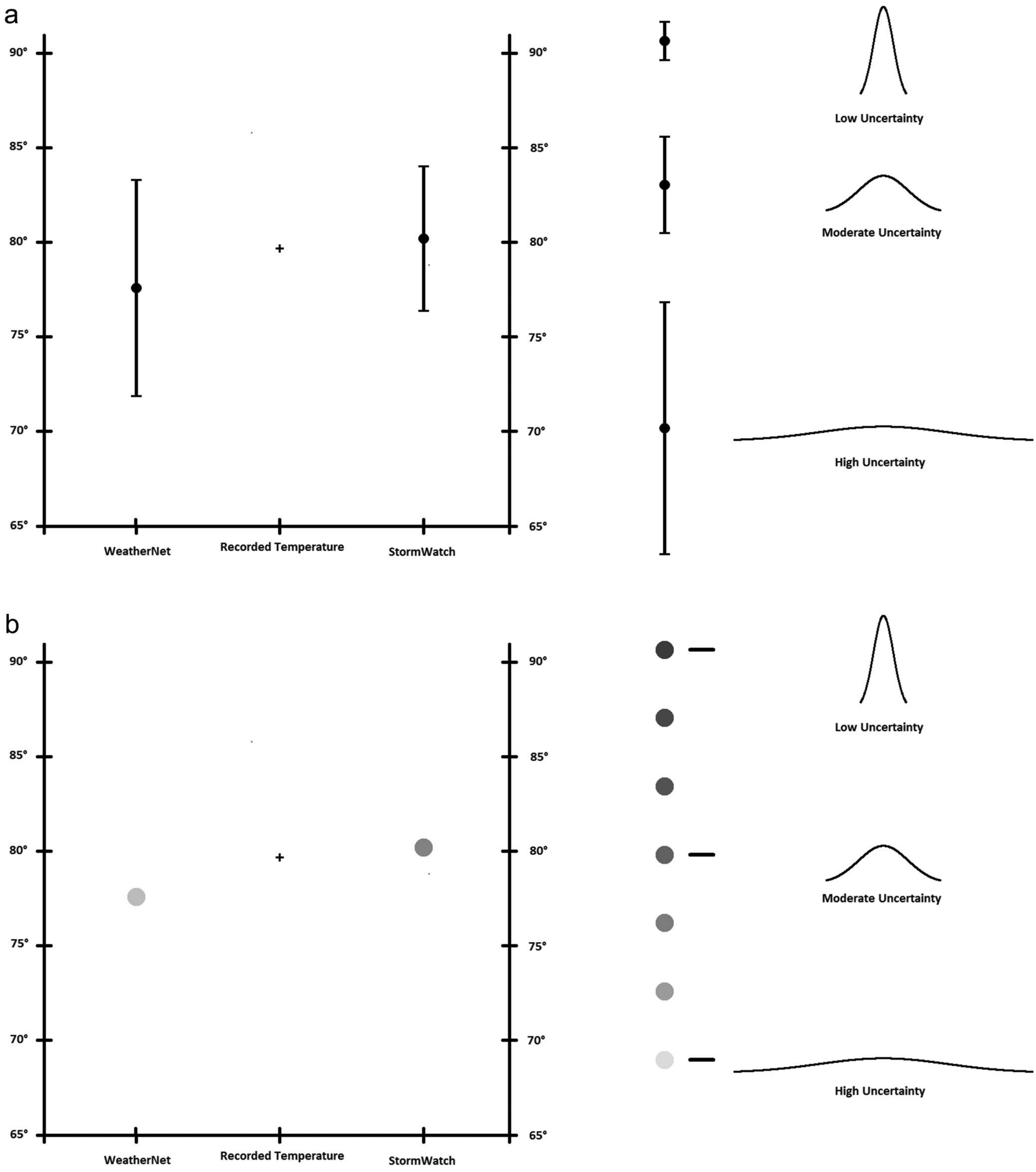

Figure 1. Examples of screen displays showing the 95% Interval (I95) glyph (top figure) and Dot glyph (bottom figure) using Distribution Set 3. These examples show one trial type with the lower standard deviation and higher mean on the right position (Stormwatch), although in the actual experiment, all glyphs were presented in equal numbers of trials in right and left positions for all four distribution sets. The right portion of the display is a key, which was presented throughout the task to provide participants with bell curve approximations of uncertainty (*SD* of underlying distributions) encodings.

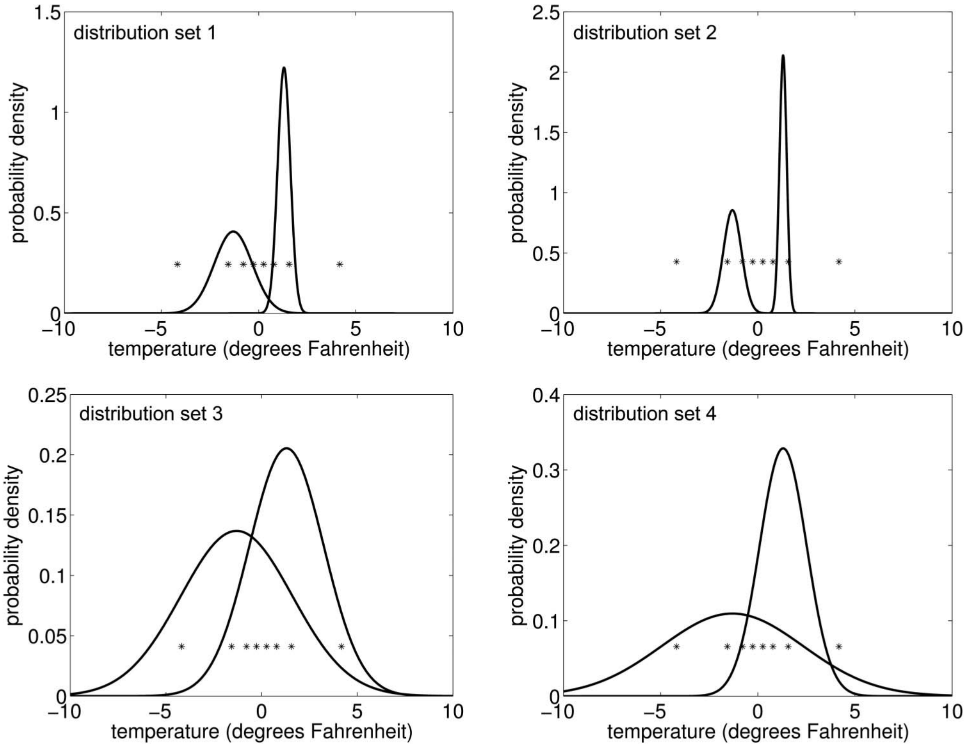

Figure 2. The four distribution sets used. The * represents the eight possible locations of "Recorded Temperature" probe points expressed in distance from the average of the two distribution means in Fahrenheit.

between-subjects factor. Distribution sets and probe points were within-subjects variables. Four distributions sets, eight probe points, and four distinct trials per probe point (as described above) resulted in a total of 128 trials per participant.

Procedure

The lights in the experiment room were turned on and the blinds shut to reduce screen glare. An experimenter ran contrast and gamma monitor tests prior to the task (http://www.lagom.nl/lcd-test/) to insure 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 midpoint of the screen. Each session lasted approximately one hour.

The experimenter read the instructions and provided a printed copy to the participant. The participants were asked to determine which forecast was more accurate given the actual temperature recorded that day (see Figure 1). They were given the following description of the scenario/task and instructions about the uncertainty visualizations presented:

Table 1
The Four Distribution Sets and Associated Standard Deviations in Fahrenheit

Distribution set	Larger *SD*	Smaller *SD*
1	0.98	0.33
2	0.47	0.19
3	2.91	1.94
4	3.64	1.21

Note. *SE* = standard deviation.

> Almost all current weather forecasts include a specific prediction for future high and low temperatures, even though the temperature may end up being different than predicted. The plots you will see in this experiment represent the outcomes of two new temperature forecasting systems for a specific date, along with the actual high temperature for that date. Both systems report forecast temperatures in a manner that indicates the amount of uncertainty in the predictions for the given day. Neither system is more accurate than the other on average over the course of the year. For each plot, you will be asked to indicate which of the two systems made the more accurate forecast, taking into account the information about the uncertainty of the forecast in your answer.
>
> The output of each forecasting system will be indicated using this graph. The right portion shows how the temperature forecasts are

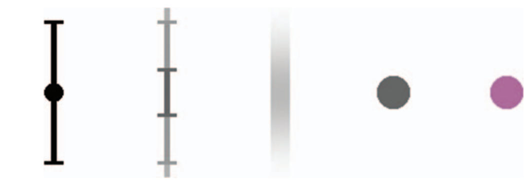

Figure 3. Examples of each glyph type used to visually present uncertainty information in temperature forecasts. From left to right: I95, IMulti, PDF, Dot, CDot.

> represented. There are examples of the forecast graphic representing three different levels of uncertainty, paired with their associated probabilistic bell curves. The left portion of the screen shows the graph that you will be using to make your decision. The higher a graphic is on the graph, the higher the temperature forecast it represents. The cross in the middle represents the actual temperature for the forecasted day. Please compare the actual temperature with the temperature forecasts to decide which forecast was more accurate.

Following completion of the task, 12 debriefing trials were conducted. These trials included a sample of the trials for each distribution set (three trials of each distribution set of varying probed temperature values). Participants were asked to make an accuracy decision, 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. Recorded responses were later transcribed into written text. These self-reports were intended to provide insight into the heuristics informing participants' decision making.

Results

Because the eight probed temperature values varied around a range of distribution means (72.23°–82.68°), we centered the means at 0 to be able to collapse across the trials for analysis. We defined a *crossover point* as the value (within the continuous range of probed temperature values from −4.18° to 4.18°) where the mean response to choose one of the forecasts crossed a threshold of .50. For each forecast pair, this threshold represented the point at which the participant changed from selecting the forecast with relatively smaller *SD* to relatively larger *SD*. The movement of the crossover point toward one distribution corresponds to a bias to choose the other distribution. This value was calculated for each individual for each of the four distribution sets, leading to four potential crossover points per individual. In some cases there was no crossover point, due to a majority of responses given to one of the forecasts at every probe point ($n = 34$, 8.0% of 424 total possible crossover points). In these cases, the data point was not included in the analysis of comparison of means, but was analyzed descriptively to assess any patterns of excluded data. There were also some cases in which multiple crossover points resulted. These data points ($n = 27$, 6.36% of 424 total possible points) were also excluded, but were not analyzed further.

First we performed a mixed-model ANOVA[1] on the crossover temperature values (in degrees Fahrenheit), with glyph as a between-subjects variable and distribution as within-subject variable, which showed no main effect of glyph, $F(4, 93) = 1.022$, $p = .40$, but a significant effect of distribution set, $F(3, 263) = 77.70$, $p < .001$ and importantly, a glyph × distribution set interaction, $F(12, 263) = 5.42$, $p < .001$. To further understand this interaction, we ran a univariate ANOVA on mean crossover points for each distribution set separately, using glyph type as a between-subjects factor, to test our main hypothesis that glyph type would influence decision making.

For Distribution Set 1, one of the two distribution sets in which the standard deviations were relatively small relative to the difference in means, the ANOVA revealed a main effect of glyph type, $F(4, 87) = 3.47$, $p = .011$, partial $\eta^2 = .137$. Planned simple contrasts compared each glyph to the 95% Interval, *I95*, ($M = .016$), showing no difference between IMulti, the multiple stepped intervals, ($M = .067$) and I95 ($SE = .14$, $p = .72$), no difference between PDF, the probability density function, ($M = .015$) and I95 ($SE = .14$, $p = .98$), but significant differences between Dot, the grayscale dot, ($M = -.34$) and I95 ($SE = .14$, $p = .012$) and CDot, the colored dot, ($M = -.29$) and I95 ($SE = .15$, $p < .042$). For Dot and Cdot, the crossover points moved negatively away from 0 (the value equidistant from the means of each distribution). This is consistent with a strategy that biased the decision toward choosing the forecast distribution with the least variability, and is opposite to the behavior that would be expected from a strategy that biases the decision toward the forecast with the highest likelihood for the temperature that actually occurred (see Figure 4a). For the three spatial-based glyphs, the crossover point falls very near 0, the point that is equidistant between the two means. This suggests that for these glyphs, participants were essentially ignoring the uncertainty information provided by the glyph and simply relying on a strategy to choose the distribution with the mean closest to the temperature value presented.

Results were similar for Distribution Set 2, the other distribution set in which the *SD*s were relatively small relative to the difference in means. The ANOVA revealed a main effect of glyph type, $F(4, 94) = 4.19$, $p = .004$, partial $\eta^2 = .151$. Planned simple contrasts compared each glyph to the standard I95 ($M = .12$), showing no difference between IMulti ($M = .06$) and I95 ($SE = .11$, $p = .57$), no difference between *PDF* ($M = .11$) and *I95* ($SE = .11$, $p = .92$), but a significant difference between *Dot* ($M = -.19$) and I95 ($SE = .11$, $p = .005$) and CDot ($M = -.21$) and I95 ($SE = .11$, $p = .004$), consistent again with a bias in responses for the dot-based glyphs toward the weather forecast distribution with the smaller *SD* (see Figure 4b). As in Distribution Set 1, the direction of the decision criterion for the dot-based glyphs was opposite to that of a decision based on the highest likelihood. This suggests that the decisions were based more on relative variability of the distributions than on relative likelihood. Also consistent with the results of Distribution Set 1, there was little use of uncertainty information for the spatial-based glyphs.

For Distribution Set 3, one of the two distribution sets in which the *SD*s were relatively large relative to the difference in means, there was also a significant effect of glyph type, $F(4, 87) = 2.97$, $p = .024$, partial $\eta^2 = .120$. The results of the planned contrasts were similar to the above distributions in finding no significant differences between the I95 ($M = -1.28$) and IMulti ($M = -1.06$) or PDF ($M = -1.35$, *p*s = .49, 80, respectively), a significant difference between Dot ($M = -.47$) and I95 ($SE = .30$, $p < .009$), and a marginal difference between CDot ($M = -.68$) and I95 ($SE = .31$, $p = .056$). However the magnitude and direction of the effect is different. All glyphs showed a bias toward choosing the forecast with the smaller *SD*, but the bias was much greater for the three spatial glyphs (I95, IMulti, PDF) compared to Dot and CDot (see Figure 4c).

For Distribution Set 4 (large *SD*s relative to the difference in means), the results were similar to those seen in Distribution Set 3, showing a main effect of glyph type, $F(4, 75) = 3.35$, $p = .014$,

[1] A linear mixed-model, in contrast to Repeated Measures ANOVA, allows for analysis of all of the data even in circumstances of missing data, as in the current data set, as some participants were missing a crossover point for some distribution set/glyph type combinations.

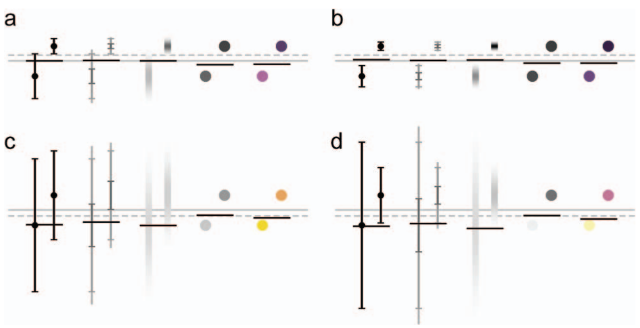

Figure 4. Mean crossover points plotted on each glyph type for (a) Distribution Set 1, (b) Distribution Set 2, (c) Distribution Set 3, and (d) Distribution Set 4. The solid light gray line represents 0, the value equidistant from the means of the two distributions. The dashed light gray line is the crossover point between the highest likelihood answer. The short black lines represent the mean crossover points for the responses to each glyph type. As the black lines move lower, this indicates a bias to choose the forecast with the smaller *SD* (pictured here on the right of each pair).

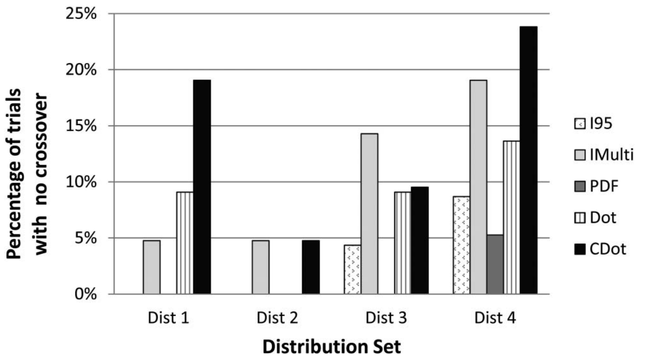

Figure 5. Percentage of trials in which there was no crossover point because of a bias to choose the less variable distribution, presented as a function of distribution set and glyph type.

partial $\eta^2 = .152$. There were no differences between I95 ($M = -1.42$) and IMulti ($M = -1.16$) or I95 and PDF ($M = -1.61$, *p*s = .45, .57, respectively). There was a significant difference between I95 and Dot ($M = -.50$; $SE = .34$, $p = .009$), but an effect only approaching significance between I95 and CDot ($M = -.80$; $SE = .38$, $p = .10$). As in Distribution Set 3, responses for all glyphs were biased toward the forecast with the smaller *SD*, but this bias was greater for the spatial glyphs compared to the dots (see Figure 4d).

For both Distribution Sets 3 and 4, the highest likelihood decision is in the negative direction (−.52), which is also consistent with a bias to choose the distribution with the lower variability (see Figures 4c and 4d). For the dot-based glyphs, given that the crossover points fall very close to the highest likelihood decision criterion, we are unable to distinguish between the likelihood and variability accounts. For the spatial glyphs, while the crossover point also moved in a negative direction, it is even more negative than the likelihood answer would predict, more likely suggesting a strategy to choose the distribution with the smaller *SD* rather than a strategy based on likelihood. Furthermore, the reduced negative bias for the dot-based glyphs compared to the spatial glyphs is consistent with the use of a variability strategy, but modulated by potentially less salient information for variability when represented by color or brightness instead of a spatial dimension. We discuss the possible strategies used in more detail below.

We also performed a descriptive analysis of "no crossover" trials. Failure to show a crossover point overwhelmingly resulted from a bias to choose the forecast with the lower *SD* (94% of "no crossover" trials). While this happened across all distribution sets and glyphs (see Figure 5), the frequency of trials with no crossovers was lowest for Distribution Set 2 (small *SD*s) and greatest for Distribution Set 4 (large *SD*s). This analysis is consistent with the ANOVA and figures presented above, suggesting that the bias to choose the less variable distribution is likely even greater than the statistical analysis reveals.

Participants' debriefing responses were closely examined on a per-trial basis to determine if any emergent strategies (sometimes referred to as "fast and frugal" heuristics) had informed decision-making during the experiment (Gigerenzer & Todd, 1999). It should be noted that the debriefing trials did not sample all trial types (as defined by the set of all actual temperatures for each forecast distribution set) presented in the experiment. Three recorded temperature values were presented for each distribution set, one of which fell within the interval between the two distribution means, and two of which fell outside of this interval on opposite ends of the distribution. Thus the descriptive conclusions that we can make about self-reported strategies are limited to the 12 debriefing trials that were presented.

Six distinct strategies were observed from the debriefing data across all five glyph types (interrater reliability as measured by *ICC* = 0.815, 95% CI [.795, .833], see Table 2). Strategy 1 involved participants choosing the forecast that had the center of its predicted temperature range closer to the actual temperature. Participants who used this strategy mostly or completely ignored uncertainty information (a *distance-to-center* heuristic). This strategy was the most commonly reported. Strategy 2 involved participants implementing a bias to choose the less variable distribution regardless of distance to the center of each distribution. In this *less-variable-distribution* heuristic, participants reported that they placed the most emphasis on variability of the distribution to inform decision making and mostly ignored relative distance of the actual temperature to the center of each forecast distribution. Strategy 3 involved participants utilizing both a distance-to-center and less-variable-distribution decision bias. In other words, participants chose the forecast distribution that was both closest to the actual temperature and less variable. Participants reported using this strategy the second most frequently.

A second set of strategies, Strategies 4, 5, and 6, was more hierarchical in nature, with participants first coming to a decision-making

Table 2
Percent of Debriefing Trials Grouped Into Six Strategy Categories

Strategy	1	2	3	4	5	6
Glyph						
I95	59.68%	5.14%	22.53%	8.30%	1.19%	3.16%
IMulti	47.44%	1.86%	35.81%	9.77%	1.86%	3.26%
PDF	54.70%	0.85%	25.21%	10.68%	8.12%	0.43%
Dot	62.81%	6.20%	13.22%	8.68%	5.79%	3.31%
CDot	46.18%	11.24%	20.08%	16.06%	6.02%	0.40%
Average	54.16%	5.06%	23.37%	10.69%	4.59%	2.11%

impasse (due to insufficient information using one set of decision criteria, such as a distance to center) and then considering secondary information to complete their decision. Strategy 4 involved participants reporting choosing the less variable forecast when the actual temperature appeared relatively equidistant to the center of each forecast. Strategy 5 involved participants choosing the forecast distribution with its center closest to the recorded temperature when each forecast distribution consisted of relatively equivalent variability. Strategy 6 involved participants choosing the more variable distribution if the actual temperature was visibly equidistant to the center of both predicted forecasts. For the trials presented, this strategy was reported least frequently.

In summary, most participants instituted “fast and frugal” heuristics such as Strategy 1, which relied on distance-to-center, Strategy 2, in which the less-variable-distribution was chosen regardless of distance, and Strategy 3, which relied on a combination of distance-to-center and less-variable-distribution heuristics. When fast and frugal heuristics were more difficult to implement, participants relied on more hierarchical heuristics that utilized a second set of information to resolve a conflict or closeness in initial decision criteria, such as when the forecast distributions appeared relatively equidistant from the actual temperature or when the forecast distributions appeared relatively equally variable.

This descriptive qualitative analysis provides insight into the information being utilized by decision-makers, with variability and closeness to center largely biasing decision-making in the 1D presentation of weather-specific uncertainty data. While these data are based on only a sample of the trials presented in the experiment, the self-reports are consistent with the calculated crossover points, which show that either variability is generally ignored, (spatial glyphs in Distribution Sets 1 and 2) or there is a bias to choose the less variable distribution (spatial glyphs in Distribution Sets 3 and 4 and generally the dot glyphs).

Discussion

The goal of this study was to examine how varying the amount of uncertainty and its visual representation would influence decisions by nonexperts in a weather forecast scenario. We set out to determine whether a more explicit visual representation of the uncertainty in weather forecasts would lead viewers to use the uncertainty differently, and whether this effect would be influenced by the amount of variability presented. The experiment revealed two main findings. The first showed that the overall level of variability displayed in each pair of distributions significantly affected the participants’ decisions. For the first two distribution sets characterized by lower standard deviations relative to the difference of means, participants’ crossover points were grouped around zero, which indicates that the point at which the participant changed from selecting one distribution to the other is equal in distance from the mean of each distribution. In these cases, it appears that participants ignored the relative size and overlap of the distribution sets and instead selected the glyph that was physically located closest to the temperature value given. However, in Distribution Sets 3 and 4, which were created with larger standard deviations, the crossover points were collectively in the negative range of values. This indicates a shift in decisions to choose the less variable (smaller *SD*) distribution.

These results suggest that in the cases where there is less variability in the distribution and less apparent overlap between each glyph in a set, participants default to using a strategy that only incorporates distance from the temperature value presented. This strategy was also reflected in the self-report data with a high percentage of responses falling into the distance-to-center strategy category. In contrast, when the glyphs in a distribution set possess larger *SD*s and appear to overlap, using a distance strategy may become more difficult. Therefore, participants may need to use other information in addition to distance or employ more complex or hierarchical decision-making strategies. Both the actual decisions and the self-report data support the use of alternate strategies that involved either the less-variable-distribution or a combination of distance-to-center and less-variable-distribution. The self-reports also suggest that a bias toward choosing the less variable prediction may have been due to participants’ different interpretation of uncertainty. Instead of understanding the *SD*s in terms of variability in predictions, participants may have interpreted them as indicating how correct or reliable the forecast was. With this interpretation, it is logical that participants would more often choose the distribution associated with less uncertainty.

The second notable finding is that of an interaction between glyph type and distribution set: participants made significantly different judgments for two separate categories of glyph type within each distribution set. One category included Dot and Cdot, which encoded uncertainty as brightness or as a combination of color and brightness. The other included PDF, I95, and IMulti, in which temperature was represented spatially, incorporating multiple encoding techniques that included size, interval bars, and brightness. For Distribution Sets 1 and 2, Dot and Cdot had crossover values that were located in a more negative range than PDF, I95, and IMulti. As mentioned above, this shift in crossover point suggests a bias in choosing the forecast with lower variability. PDF, I95, and IMulti grouped closely around zero, suggesting that these glyphs provoked the distance strategy to a greater degree than Dot and Cdot. For Distribution Sets 3 and 4, participants also made significantly different judgments between these two glyph groupings. Both groups’ values were located in the negative range. However, PDF, I95, and IMulti were significantly more negative, showing an even greater bias to choose the less variable distribution. As mentioned above, if in fact a variability strategy was used in all of the glyph types, the reduced bias to choose the less variable distribution in the nonspatial glyphs could have been due to color and brightness serving as less salient cues to variability than the spatial representations.

The purpose of creating three different glyphs to represent the distribution in a spatial dimension was to investigate whether providing more information about the underlying distribution than a typical 95% confidence interval would lead people to make a decision that approached the maximum likelihood criteria more closely. Specifically, we presented multiple intervals and open bars in IMulti to attempt to convey a normal distribution without discrete endpoints, and we presented the gradient in PDF to explicitly encode the probability density function. We did not, however, explain the statistical meaning of the information presented. Notably, neither of these glyphs differed significantly from the more familiar 95% interval (I95). This result suggests that either the visual information presented in the glyphs was not interpreted correctly or it was ignored. Particularly for the distribution sets with overall lower *SD* (sets 1 and 2), the result of a crossover point clearly at zero suggests the tendency to choose a distance-to-center strategy over the use of uncertainty. There was no evidence that this strategy changed with the implementation of IMulti or PDF. Future work might assess whether more explicit instruction or training about how the underlying distributions are represented in the glyphs changes decision-making behavior. If our hypothesis is correct, that participants chose the less variable forecast because they interpreted the lower variability in prediction as a more reliable or accurate forecast, then it is possible that instructions that distinguish between these two interpretations of uncertainty would change behavior. For distribution sets 3 and 4, while the crossover point moved negatively away from zero, there still were no significant differences among the three spatial glyphs. This suggests that although a different strategy (likely the less-variable-distribution heuristic) was used for these distribution sets, it was used similarly across all three glyphs. It is also possible that the lack of difference between the spatial glyphs is not an issue with understanding the glyphs themselves, but a more fundamental problem with reasoning about a probability density function. Future work using explicit training in reasoning with uncertainty could help to address this possibility.

Comparison of our results with the highest likelihood answer, defined by the value of the probability density function at the presented temperature, sheds additional light on participant decision-making strategy. For Distribution Sets 1 and 2, the highest likelihood answer[2] moves in the positive direction (+.52) relative to the value equidistant from both means (0), opposite to the direction of the performance seen with the dot-based glyphs. For Distributions Sets 3 and 4, the highest likelihood moves in the negative direction (−.52), consistent with the direction of responses found in our study. However, this is also the same direction as predicted by the less-variable-distribution strategy. Thus, while converging evidence from the debriefing trials suggests that participants were probably not using higher likelihood to make their decisions for the dot-based glyphs, we cannot distinguish use of likelihood from the less-variable-distribution strategy because of the nature of the distributions in sets 3 and 4.

These findings have important implications for the use of uncertainty visualizations in applied domains. It is possible that the underlying distribution matters more in guiding uncertainty-based decision making when the underlying distribution of data is both inherently more uncertain and widely distributed. In these cases in the current study, participants were biased toward the distribution with lower *SD*. In application domains where greater amounts of uncertainty are presented, visualization designers should be aware of this bias and further research is needed to define possible glyph types or training that would reduce this heuristic. Related to glyph-type effects on decision making, dot-based visualizations may have guided participants' decision making differently in the face of large variability by mostly eliminating size-based spatial cues (which were very salient in the PDF, I95, and IMulti glyphs) to the underlying variability. Our findings suggest that when selecting a glyph type, designers should consider the distribution of uncertainty as a primary factor that can bias the way the viewer may interpret the visualization. However, more work is needed to further examine how the characteristics of the uncertainty distributions interact (such as the overall amount of variability and the relative variability among different distributions) to influence the types of heuristics used.

Caution must be taken in generalizing these findings to include uses of uncertainty visualizations in other domains and tasks. One piece of the experiment space left unexplored by this experiment is domain-specific effects on decision making under uncertainty. First, participants often expect weather-forecasting models to include uncertainty information, due to first-hand experience with inaccurate forecasting (Joslyn & Savelli, 2010). Perhaps presenting uncertainty information in a domain where uncertainty is not expected would lead to greater trust of the data and a better understanding of the underlying distributions represented by each glyph, less skewed by weather-specific expectations. In addition, weather forecasting itself consists of more specific subdomains aside from temperature. For example, it remains unclear if glyphs represented within 2D space—as is required by some hurricane forecasting uncertainty visualizations—would demonstrate similar results and/or benefit from verbal descriptions of uncertainty (Cox et al., 2013). It is also the case that generally people are more experienced with uncertainty included in precipitation forecasts compared to temperature forecasts. Future work could examine the influence of the subdomain of weather on use and interpretation of uncertainty information.

Second, we asked participants to judge which forecast was more accurate on a given day, but a host of other tasks or questions could have been performed using uncertainty visualizations. An example of another methodology includes asking nonexperts to make judgments or decisions with consequences. Joslyn and LeClerc (2012), for instance, asked decision makers to make a judgment of whether or not to salt a road based on a series of deterministic or probabilistic forecasts. Perhaps framing the presentation of weather-based visualizations in the context of a real-world decision would both better illuminate the probabilistic nature of uncertainty and get participants to more carefully consider their decision (lessening the use of hard and fast heuristics). Other uncertainty visualization work has collected nonexperts quantitative and qualitative evaluations of uncertainty information in visualizations by tasking nonexperts with nonconsequential search and counting tasks unrelated to a specific domain- (Sanyal et al., 2009; MacEachren et al., 2012).

In summary, we used glyphs representing single scalar values to investigate decision-making strategies employed by nonexpert users in a realistic weather forecast context. We found that users frequently employ strategies that ignore uncertainty-based information or show decisions that are biased toward lower variability distributions, and that these strategies sometimes lead to decision making that is not consistent with a highest likelihood solution. In addition, we found that decision-making performance varied by glyph type, with different strategies and decisions employed for spatial versus nonspatial encoding of uncertainty, but that the additional information about the distributions provided in the variations of the spatial glyphs did not change performance. We hope that knowledge gained through this investigation can be used in weather forecasting, emergency preparedness, and other fields to improve communication of uncertain information to nonexpert users.

[2] In addition to the highest likelihood transition point between the two distributions considered here, there is a second transition outside of the interval between the two means and associated with relatively low values in both PDF functions. We did not consider this second transition point in the analysis of the results.

References

Aerts, J. C. J. H., Clarke, K. C., & Keuper, A. D. (2003). Testing popular visualization techniques for representing model uncertainty. *Cartography and Geographic Information Science, 30,* 249–261. http://dx.doi.org/10.1559/152304003100011180

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396. http://dx.doi.org/10.1037/1082-989X.10.4.389

Bisantz, A. M., Marsiglio, S. S., & Munch, J. (2005). Displaying uncertainty: Investigating the effects of display format and specificity. *Human Factors, 47,* 777–796. http://dx.doi.org/10.1518/001872005775570916

Borland, D., & Taylor, R. M. (2007). Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications, 27,* 14–17. http://dx.doi.org/10.1109/MCG.2007.323435

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,* 651–667. http://dx.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,* 143–156. http://dx.doi.org/10.1615/Int.J.UncertaintyQuantification.2012003966

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics, 29,* 89–93. http://dx.doi.org/10.1111/j.1467-9639.2007.00267.x

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180. http://dx.doi.org/10.1037/0003-066X.60.2.170

Evans, B. J. (1997). Dynamic display of spatial data-reliability: Does it benefit the map user? *Computers & Geosciences, 23,* 409–422. http://dx.doi.org/10.1016/S0098-3004(97)00011-3

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace *p* values. *Zeitschrift für Psychologie/Journal of Psychology, 217,* 27–37.

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science, 22,* 392–399. http://dx.doi.org/10.1177/0963721413491570

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 3–36). New York, NY: Oxford University Press.

Hengl, T. (2003). Visualisation of uncertainty using the HIS colour model: Computations with colours. *Proceedings of the 7th International Conference on GeoComputation* (pp. 8–17). Southampton, UK: University of Southampton.

Howard, D., & MacEachren, A. M. (1996). Interface design for geographic visualization: Tools for representing reliability. *Cartography and Geographic Information Science, 23,* 59–77. http://dx.doi.org/10.1559/152304096782562109

Jiang, B., Ormeling, F., & Kainz, W. (1995, March). Visualization support for fuzzy spatial analysis. In *Proceedings of ACSM/ASPRS conference, Charlotte, NC* (pp. 291–300).

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied, 18,* 126–140. http://dx.doi.org/10.1037/a0025185

Joslyn, S., Nemec, L., & Savelli, S. (2013). The benefits and challenges of predictive interval forecasts and verification graphics for end users. *Weather, Climate, and Society, 5,* 133–147. http://dx.doi.org/10.1175/WCAS-D-12-00007.1

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications, 17,* 180–195. http://dx.doi.org/10.1002/met.190

Joslyn, S., Savelli, S., & Nadav-Greenberg, L. (2011). Reducing probabilistic weather forecasts to the worst-case scenario: Anchoring effects. *Journal of Experimental Psychology: Applied, 17,* 342–353. http://dx.doi.org/10.1037/a0025901

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228. http://dx.doi.org/10.1037/0033-295X.94.2.211

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,* 139–160. http://dx.doi.org/10.1559/1523040054738936

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,* 2496–2505. http://dx.doi.org/10.1109/TVCG.2012.279

Marzouk, Y. M., Najm, H. N., & Rahn, L. A. (2007). Stochastic spectral methods for efficient Bayesian solution of inverse problems. *Journal of Computational Physics, 224,* 560–586. http://dx.doi.org/10.1016/j.jcp.2006.10.010

Masalonis, A. J., & Parasuraman, R. (2003). Effects of situation-specific reliability on trust and usage of automated air traffic control decision aids. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 47,* 533–537.

Morone, A., & Ozdemir, O. (2012). Displaying uncertain information about probability: Experimental evidence. *Bulletin of Economic Research, 64,* 157–171. http://dx.doi.org/10.1111/j.1467-8586.2010.00380.x

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review, 19,* 601–607. http://dx.doi.org/10.3758/s13423-012-0247-5

Pang, A. T., Wittenbrink, C. M., & Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer, 13,* 370–390. http://dx.doi.org/10.1007/s003710050111

Rhodes, P. J., Laramee, R. S., Bergeron, R. D., & Sparr, T. M. (2003). Uncertainty visualization methods in isosurface rendering. In M. Chover, H. Hagen, & D. Tost (Eds.), *Eurographics 2003* (pp. 83–88). Goslar, Germany: Eurographics Association.

Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., & Moorhead, R. J. (2009). A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions on Visualization and Computer Graphics, 15,* 1209–1218. http://dx.doi.org/10.1109/TVCG.2009.114

Sanyal, J., Zhang, S., Dyer, J., Mercer, A., Amburn, P., & Moorhead, R. J. (2010). Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics, 16,* 1421–1430. http://dx.doi.org/10.1109/TVCG.2010.181

Schweizer, D. M., & Goodchild, M. F. (1992). Data quality and choropleth maps: An experiment with the use of color. In *G. I. S. LIS-International conference* (Vol. 2, pp. 686–699). Bethesda, MD: American Society for Photogrammetry and Remote Sensing.

Wittenbrink, C. M., Pang, A., & Lodha, S. K. (1996). Glyphs for visualizing uncertainty in vector fields. *IEEE Transactions on Visualization and Computer Graphics, 2,* 266–279. http://dx.doi.org/10.1109/2945.537309

Zuk, T., & Carpendale, S. (2007, January). Visualization of uncertainty and reasoning. *Smart Graphics: Lecture Notes in Computer Science, 4569,* 164–177. http://dx.doi.org/10.1007/978-3-540-73214-3_15

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment, 19,* 116–138. http://dx.doi.org/10.1080/10627197.2014.903653

Received May 30, 2014
Revision received September 13, 2014
Accepted October 4, 2014 ■