Public Understanding of Visual Representations of Uncertainty in Temperature Forecasts

Susanne Tak, TNO and Freudenthal Institute for Science and Mathematics Education, Utrecht University, **Alexander Toet**, TNO, and **Jan van Erp**, TNO and University of Twente

Multiday weather forecasts often include graphical representations of uncertainty. However, visual representations of probabilistic events are often misinterpreted by the general public. Although various uncertainty visualizations are now in use, the parameters that determine their successful deployment are still unknown. At the same time, a correct understanding of possible weather forecast outcomes will enable the public to make better decisions and will increase their trust in these predictions. We investigated the effects of the visual form and width of temperature forecast visualizations with uncertainty on estimates of the probability that the temperature could exceed a given value. The results suggest that people apply an internal model of the uncertainty distribution that closely resembles a normal distribution, which confirms previous findings. Also, the visualization type appears to affect the applied internal model, in particular the probability estimates of values outside the depicted uncertainty range. Furthermore, we find that perceived uncertainty does not necessarily map linearly to visual features, as identical relative positions to the range are being judged differently depending on the width of the uncertainty range. Finally, the internal model of the uncertainty distribution is related to participants' numeracy. We include some implications for makers or designers of uncertainty visualizations.

Keywords: visualization, topics, laboratory study, methods, user evaluation study, uncertainty, topics

Address correspondence to Susanne Tak, Freudenthal Institute for Science and Mathematics Education, Utrecht University, P.O. Box 85170, 3508 AD Utrecht, the Netherlands, susannetak@gmail.com.

Journal of Cognitive Engineering and Decision Making
2015, Volume 9, Number 3, September 2015, pp. 241–262
DOI: 10.1177/1555343415591275

INTRODUCTION

The quality of weather forecasts has improved dramatically in recent decades (Richardson et al., 2011). This improvement is mainly due to the maturing of numerical weather forecast models and the improved global observation of the atmosphere, particularly from satellites. But because there are fundamental physical limits to our ability to predict complex systems, weather forecasts will always contain uncertainty (Kootval, 2008). To overcome the limitations of deterministic weather forecasting in the face of uncertainty and chaos, meteorologists developed ensemble prediction methods (Ehrendorfer, 1997; Leith, 1974), which have become the state-of-the-art technique for weather forecasting (Craig, 2012; Gneiting & Raftery, 2005). In contrast to deterministic modeling techniques that produce a single forecast, ensemble prediction methods generate a collection of predictions of the same physical phenomenon using different parameter values, boundary or initial conditions, instances of stochastic phenomena, phenomenological models, numerical regimes or parameters, or combinations thereof (Craig, 2012). Such an ensemble of predictions can be converted into a probability distribution function that provides a way to infer the (un)certainty of a given future outcome.

Weather forecasts of 5, 10, or even 30 days in advance are quite common nowadays. The uncertainty in these forecasts is often presented together with the forecast itself (see Figures 1-3 for some examples of temperature forecasts that include depictions of uncertainty). The idea is that users may appreciate this uncertainty visualization ("nice to know") and may use it to reach a decision ("need to know"). It has indeed been observed that everyday users infer uncertainty

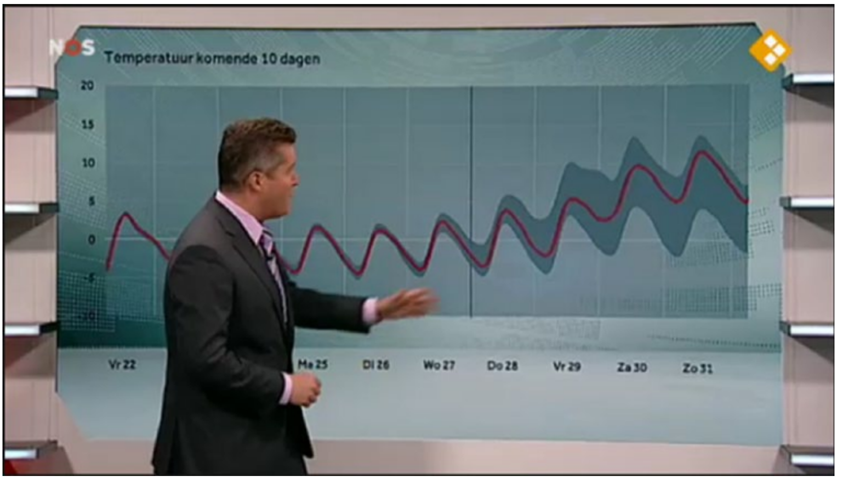

Figure 1. A visual representation of temperature forecast uncertainty in the news (screenshot taken from Dutch television station NOS, news edition, March 22, 2013).

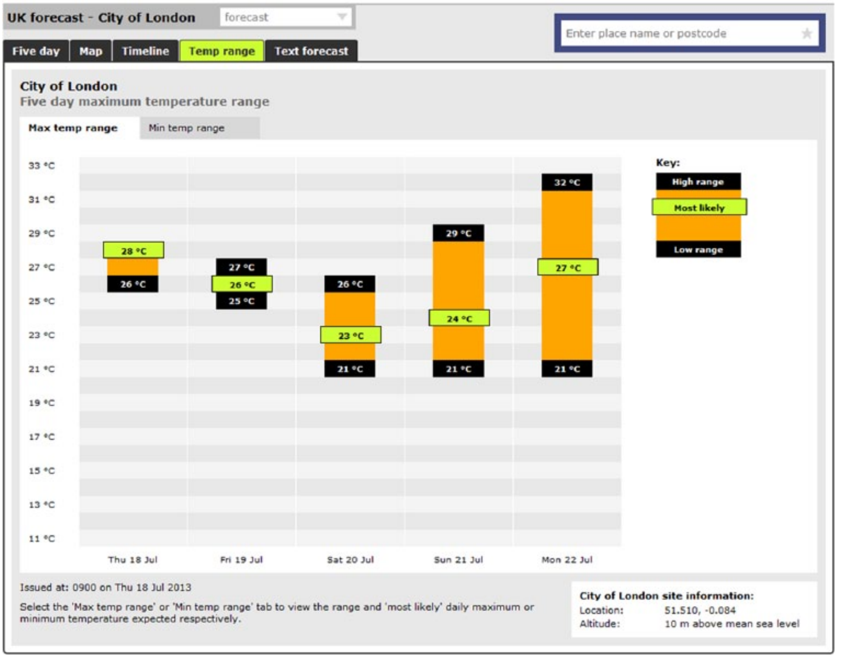

Figure 2. A visual representation of temperature forecast uncertainty (from www.metoffice.gov.uk, July 18, 2013).

into deterministic forecasts anyway and prefer forecasts that explicitly express this uncertainty (Joslyn & Savelli, 2010; Morss, Demuth, & Lazo, 2008; Morss, Lazo, & Demuth, 2010; Peachy, Schultz, Morrs, Roebber, & Wood, 2013; Savelli & Joslyn, 2013). Research has also shown that including uncertainty estimates in probabilistic weather (and hydrological) forecasts increases trust and gives people a better understanding of the possible outcomes and the amount of uncertainty in the given situation, allowing them to make better decisions (Joslyn

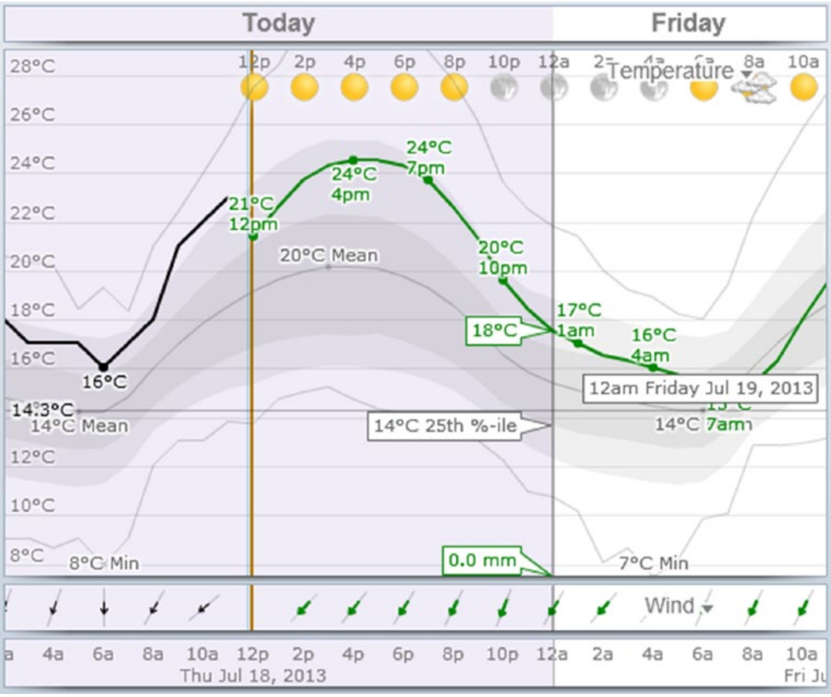

Figure 3. A visual representation of temperature forecast uncertainty (from www.weatherspark.com, July 18, 2013).

& LeClerc, 2012, 2013; Joslyn, Nemec, & Savelli, 2013; Marimo, Kaplan, Mylne, & Sharpe, 2012; Nadav-Greenberg & Joslyn, 2009; Ramos, van Andel, & Pappenberger, 2013; Roulston, Bolton, Kleit, & Sears-Collins, 2006; Roulston & Kaplan, 2009; Savelli & Joslyn, 2013). Carefully designed visual representations can indeed successfully convey uncertainty information to both experts and nonexperts (Nadav-Greenberg, Joslyn, & Taing, 2008; Roulston & Kaplan, 2009). However, the advantage of the availability of uncertainty estimates depends critically on how they are communicated (Ibrekk & Morgan, 1987; Nadav-Greenberg et al., 2008). Effectively communicating forecast uncertainty to nonexperts still remains a challenge (Demuth, Morss, Lazo, & Stewart, 2007; Joslyn & Savelli, 2010; Joslyn, Savelli, & Nadav-Greenberg, 2011; Joslyn et al., 2013; Morss et al., 2008; Savelli & Joslyn, 2009, 2013; Stephens, Edwards, & Demeritt, 2012).

In the current study, we investigate how users interpret the uncertainty visualization in temperature forecasts. Because we encounter probabilistic weather forecasts on a regular basis nowadays, this issue may seem trivial. However, visualizations are often shown without any explanation of the meaning of the uncertainty range (e.g., Figure 1). In case an explanation is provided, it is often ambiguous (e.g., in Figure 2, where the exact meaning of "high range" and "low range" is not given) or needs a certain level of expertise to comprehend (e.g., the meaning of a "25th %-ile" as provided in Figure 3). Because previous work has shown that even researchers (though generally quite familiar with uncertainty visualizations) frequently misunderstand visual representations of confidence intervals and errors bars (Belia, Fidler, Williams, & Cumming, 2005), one can expect that laymen will have the same problems.

Visualizations like those depicted in Figure 1 do not provide direct information about, for example, the probability of reaching a temperature of 10 degrees on a specific day. Tak, Toet, and van Erp (2014) suggested that users may apply an internal, cognitive model to translate these kind of visualizations into point probabilities. To test this hypothesis, they used a case from the geomodeling domain. In their experiment, participants judged the probability that the boundary between two earth layers (represented as a 1D graph plus an uncertainty range of uniform width) passed through a given location indicated by a red dot, which was presented randomly at one of nine vertical positions relative to the center line (similar to the procedure used in the present study; see Method section). They tested seven different visual representations for both the borders (e.g., solid lines, dashed lines, or a gradient boundary) and the interior (e.g., solid fill, gradient fill, parallel lines, or random lines) of the uncertainty range visualization. They found that the user's hypothetical internal model is best fitted by a normal distribution (better than, for instance, triangular, quadratic, or cubic distributions), with the highest probability corresponding to values on the center line of the uncertainty range and increasingly lower probabilities corresponding to values located at larger distances from the center line. The internal model has a wider distribution than would have been expected if the range would depict a 95% confidence interval as is common practice (i.e., it is overdispersed). Note that no values for the uncertainty ranges were provided in their experiment. In addition, Tak et al. (2014) also found that the visual depiction of the uncertainty range (i.e., the visualization type) affects these perceived probabilities.

To increase the generalizability of the previous findings by Tak et al. (2014) and to increase our knowledge about the (public) understanding of probabilistic forecasts in particular, we investigate the following research questions:

1. What does the internal model for range probability look like?
2. Does the internal model for range probability depend on the visualization type?
3. Do the perceived range probabilities depend on the width of the uncertainty visualization?
4. Does the internal model for range probability depend on numeracy?

Our first hypothesis (H1) is that people will apply a normal distribution to infer range probability from visualizations of temperature forecasts with uncertainty ranges when no further information is available. Our second hypothesis (H2) is that the way people infer range probability from these visualizations depends on the graphical representation of the uncertainty range when no further information is available. Our third hypothesis (H3) is that perceived range probabilities are independent of the width of the uncertainty visualization, such that the probability of exceeding temperature values with equal relative difference from the center value of the probability range are judged the same for wide and for narrow uncertainty ranges. Finally, our fourth hypothesis (H4) is that the shape of the applied distribution depends on people's level of numeracy. We base this hypothesis on earlier evidence that there may be a relationship between numeracy and the subjective assessment of uncertainty visualizations (Rinne & Mazzocco, 2013; Tak et al., 2014).

RELATED WORK

Uncertainty communication methods, whether they are verbal (Juanchich, Teigen, & Gourdon, 2013; Teigen & Filkuková, 2013; Teigen, Juanchich, & Filkuková, 2013; Teigen, Juanchich, & Riege, 2013), numerical (Rinne & Mazzocco, 2013), or graphical (Belia et al., 2005; Hildon, Allwood, & Black, 2012; Kootval, 2008; Lem, Onghena, Verschaffel, & Van Dooren, 2013; Spiegelhalter, Pearson, & Short, 2011), typically give rise to a range of misunderstandings. At the same time, providing uncertainty information may help to disambiguate the concept of uncertainty, increase user trust, and eventually improve decision making (Joslyn et al., 2013), particularly for people with low numeracy (Garcia-Retamero & Galesic, 2010; Garcia-Retamero & Hoffrage, 2013). In this section, we briefly review the problems that occur in understanding probabilistic forecasts in general, the various visualization approaches that have been developed to communicate uncertainty in forecasts, and the specific problems that occur with the interpretation of these uncertainty visualizations.

Understanding Probabilistic Forecasts

Even experts (including meteorologists; de Elía & Laprise, 2005; Murphy, Lichtenstein, Fischhoff, & Winkler, 1980) sometimes misunderstand probabilities and misinterpret probabilistic forecasts (Tversky & Kahneman, 1974).

Although nonexperts can interpret probability of weather events in a general sense (e.g., “70% is higher than 30%”), the specific meaning of the forecasts remains unclear to many (Gigerenzer, Hertwig, Van Den Broek, Fasolo, & Katsikopoulos, 2005; Joslyn & Savelli, 2010; Morss et al., 2008; Murphy et al., 1980). In addition, people appear to have a strong tendency to convert probabilistic forecasts into deterministic ones; that is, they make “deterministic construal errors” (Savelli & Joslyn, 2013). For instance, it has been observed that the upper and lower bounds in predictive temperature interval visualizations are sometimes misinterpreted as diurnal temperature fluctuations (Savelli & Joslyn, 2013). Predictive interval visualizations may therefore require a more direct depiction of uncertainty that blocks automatic deterministic interpretations (Savelli & Joslyn, 2013).

Techniques for Visualizing Uncertainty

There are roughly three different ways to visualize uncertainty: by varying the graphical properties of the visualization (intrinsic uncertainty representation; Gershon, 1998), by adding uncertainty information to the visualization (extrinsic uncertainty representation; Gershon, 1998), and by animating the visualization (Ehlschlaeger, Shortridge, & Goodchild, 1997).

The first approach deploys various techniques to vary the graphical properties of depicted entities, such as blur, sketchiness, transparency, size, texture, and color saturation (Boukhelifa, Bezerianos, Isenberg, & Fekete, 2012; Finger & Bisantz, 2002; Griethe & Schumann, 2005; for a recent taxonomy, see Potter, Rosen, & Johnson, 2012). For example, blurring or degradation (i.e., reducing the saliency; Bisantz et al., 2009) of the data is intuitively related to uncertainty: The harder it is to see or recognize something, the more uncertain it appears (Finger & Bisantz, 2002). However, blurring or degradation can also inadvertently be interpreted as poor visualization quality (Riveiro, 2007).

The second approach is to add uncertainty information to a visualization, such as glyphs (graphical elements that can convey a number of variables through variations in their size, shape, orientation, texture, and color), geometry, labels, and icons. For example, positional uncertainty can be indicated by overlaying a glyph, the size of which becomes larger the more uncertain the location is (Andre & Cutler, 1998; Brolese & Huf, 2006; Pang, Wittenbrink, & Lodha, 1997). Geometric techniques include contour lines and isosurfaces (Pang et al., 1997; Pöthkow & Hege, 2011). Also, textual or numerical information about the magnitude of uncertainty can be added to the visualization (Kootval, 2008; Nadav-Greenberg & Joslyn, 2009). Adding graphical representations of uncertainty information to a data visualization may result in data obscuration and user distraction (Cedilnik & Rheingans, 2000). Although uncertainty annotations have been designed that minimize these effects (Cedilnik & Rheingans, 2000), the added amount of information can still negatively influence the user’s response time, as it sometimes requires more cognitive processing (Andre & Cutler, 1998).

Finally, uncertainty information can also be visualized through animation. However, this may lead to annoying and distracting blinking and flicker effects (Evans, 1997) and may result in cluttered or overcrowded displays.

In weather forecasting in particular, there are various ways to communicate the results of an ensemble prediction in graphical form such as spaghetti plots (plots including multiple lines, each representing the outcome of one model prediction), line graphs with error bars, contour box plots, bar charts, fan charts (line graphs with an uncertainty “band”), threshold maps (using color or gray scales), and summary tables (Demeritt et al., 2007; Demeritt, Nobert, Cloke, & Pappenberger, 2013; Kootval, 2008; Pappenberger et al., 2013; Sanyal et al., 2010; Whitaker, Mirzargar, & Kirby, 2013). The study reported in this paper investigates seven different visual formats of hypothetical ensemble temperature predictions to assess their effectiveness in communicating prediction uncertainty.

Understanding Uncertainty Visualizations

Because many people (even experts) frequently misunderstand visual representations of uncertainty (Belia et al., 2005), there is a need for better graphical conventions that unambiguously convey the notion of probability

(Cumming, 2007). Research on visual uncertainty communication mostly focuses on the development of new graphical uncertainty representations, with little attempt to evaluate their effectiveness for users (Bisantz et al., 2009). Even less research has been done into how well uncertainty visualization supports decision making (Joslyn et al., 2013; Savelli & Joslyn, 2013; Zuk & Carpendale, 2007). It is often simply taken for granted that visual depictions of uncertainty will be useful for decision making (MacEachren et al., 2005). As a result we still do not have a comprehensive understanding of the parameters that influence successful uncertainty visualization (MacEachren et al., 2005).

When no information on the nature of the probability distribution is available, people sometimes assume a uniform probability distribution, both for graphical (Ibrekk & Morgan, 1987) and numerical (Rinne & Mazzocco, 2013) uncertainty representations.

A notorious example of the misunderstanding of a graphical uncertainty representation is the deterministic construal error induced by the well-known "Cone of Uncertainty" graphic used by both the U.S. National Hurricane Center and the media to communicate hurricane risk to the public prior to landfall (Broad, Leiserowitz, Weinkle, & Steketee, 2007). The main elements of this graphic are a black line representing the predicted path of the hurricane center, centered on a white "cone" representing the potential geographic range of the track. Despite the attempt of the forecast community to make a user-friendly product, this type of hurricane-warning graphics is misinterpreted by a large part of the public. Although the track line only represents the predicted (potential) track of the a hurricane center, the lay public typically fails to appreciate both the uncertainty about it or the statistical meaning of the wider "cone" of uncertainty about its projected course. The white cone is often incorrectly interpreted as the extent of the hurricane, its intensity, or the potential swath of destruction (Broad et al., 2007). As a result, people often fail to understand that the hurricane will potentially affect a much larger area than just the cone depicting the uncertainty about the track of the eye of the storm. People wrongly assume that only areas along the track line are at risk (over distances up to the boundary of the cone), whereas areas outside the cone will not be impacted. Another source of confusion is the fact that the white cone has been obtained by thresholding the actual overall spatial probability distribution (i.e., the continuous distribution was converted into a binary one by replacing all values above a given threshold with a predefined maximum value and all values below the threshold with predefined minimum value), resulting in a loss of information (the variance in probabilities over the white area is no longer available). It has been observed that this may cause an overestimation of the probability along the center line, even by experts (Kirlik, 2007).

A previous study on the perception of point probability in graphs with visual uncertainty bands (Tak et al., 2014) found that observers (when given no explanation of the "mathematical" meaning of the uncertainty band) intuitively apply an internal model of uncertainty bands that closely resembles a normal distribution. The current study extends this previous work by investigating the effects of type and overall width of ensemble prediction visualizations (presented without any additional information) on range probability estimates.

METHOD

We follow up on the findings by Tak et al. (2014) with three key differences. First, we examine range probability as opposed to point probability, as a range probability question (e.g., "What are the chances that it will be warmer than 25 degrees Celsius tomorrow?") is more realistic than a point estimate (e.g., "What are the chances that it will be 25 degrees Celsius tomorrow?") in the domain of temperature forecasts. Second, Tak et al. (2014) used uncertainty ranges with uniform widths, which are unlikely for temperature forecasts over time. Therefore, in the current study, we examine the effect of varying (i.e., monotonously increasing) the width of the uncertainty range. Third, the visualizations in Tak et al. (2014) concerned an unfamiliar, abstract domain (the location of the boundary between two earth layers). This means that users will not have an a priori internal model (built on experience), whereas this may be the case with respect to weather forecasting.

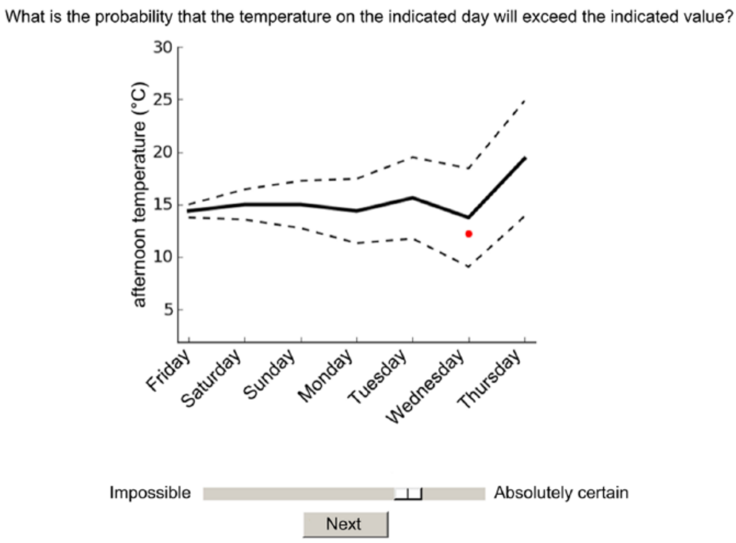

Figure 4. Screen shot of a stimulus in the *dashed border* condition (Dutch instructions and labels were used in the experiment).

Design and Procedure

Before starting the actual experiment, the participants first viewed some slides presenting a "cover story" to provide some context about uncertainty visualizations in temperature forecasts and explaining the experimental procedure. The cover story and the instructions were as follows (translated from Dutch):

> You are about to participate in an experiment on uncertainty visualizations. This experiment will last about 15 minutes. Nowadays weather forecasts increasingly include graphical representations of the uncertainty of the given predictions. In this experiment you will see a series of temperature graphs provided by different weather forecasters. Each of these graphs represents the predicted temperature over a 7-day period together with the associated uncertainty. [Note: At this time, the participant was presented with an example graph like the one in Figure 4 but without the red dot] It is your task to estimate the probability that the afternoon temperature on a given day will exceed a given threshold value, indicated by a red dot in the graph. [Note: At this time, the participant was presented with an example graph including a red dot, like Figure 4] You can enter your response by dragging the slider to the position corresponding to your estimate. [Note: At this time, the participant was shown a linear scale with endpoints labeled, respectively, *impossible* to *absolutely certain* and a slider that could be moved along this scale using the mouse pointer.] By clicking on the "Next" button, you can indicate that you understand these instructions and the actual experiment will start.

In this experiment, participants were asked to estimate a range probability by judging the probability that the afternoon temperature on a given day would exceed the temperature indicated by a red dot in a given uncertainty visualization (Figure 4). The different visualizations and the possible positions of the red dot are described in detail in the next section. Participants reported their probability estimates by positioning a slider on a continuous scale with end points labeled *impossible* (translated to 0 in our results) to *absolutely certain* (translated to 100 in our results). Similar to Tak et al. (2014), no numerical markers were added along this scale. The rationale was that labeling could prime participants to form a linear relation between their probability judgment and the uncertainty visualization (i.e., to assign a value of 50% to the middle of the scale, 25% and 75% to the first and last quarter points, and so forth), whereas previous work suggests that such a linear relationship between visual uncertainty features and perceived certainty does not necessarily exist (Sanyal et al., 2009; Tak et al., 2014).

After completing the experiment, the participants were asked (using an open question) to describe how they had assessed the stimuli (e.g., if they had ever judged the probability that a temperature would be higher than indicated by the red dot to be *impossible* or *absolute certain* at any point, and if so, why). Also, participants' numeracy was assessed using the Subjective Numeracy Scale (Fagerlin et al., 2007). Finally, we asked participants to report their highest level of education completed (seven categories: 1 = *primary education/no education*; 2 = *lower vocational education*; 3 = *lower secondary education*; 4 = *higher secondary education*; 5 = *BSc*; 6 = *MSc*; 7 = *PhD*).

Similar to the study by Tak et al. (2014), the experiment used a mixed design, with visualization type as a between-subject independent variable and position and width as within-subject independent variables. Participants were randomly assigned to one of seven groups, corresponding to the seven visualization types shown in Figure 5.

Stimuli

Figure 5 shows the seven visualization types that were used in this study. These visualizations are similar to the ones used by Tak et al. (2014) in their study on the visualization of the positional uncertainty of the boundary between two earth layers (represented as a 1D graph plus an uncertainty interval with a uniform width). All seven visualizations represented seven data points connected by a continuous black line (we will refer to this line as the center line). The seven data points represented the predicted temperature values for 7 days ahead. The visualizations differed in the graphical representation of the uncertainty range, which was always symmetrical around the center line.

The width of the uncertainty range increased monotonously along the horizontal (temporal) axis (in contrast to the study by Tak et al., 2014, who used an uncertainty range with a uniform width). The solid border and dashed border visualizations both represent the edges of the uncertainty range by a (respectively, solid or dashed) line. The band and gradient visualizations fill the area between the edges of the uncertainty range with, respectively, a uniform or a gradient grayscale distribution. The thinning lines and random lines visualizations fill the area between the edges of the uncertainty range with lines, either parallel and with a decreasing density from the center outward (thinning lines) or with (semi) randomly generated lines (random lines). The random lines representation resembles a type of uncertainty visualization that is sometimes also called an ensemble prediction, (model) realizations, or more informally, a spaghetti plot. Finally, the error bars visualization uses "traditional" error bars.

In the course of the experiment, the participants were asked to judge the probability that the afternoon temperature on one of two days (Sunday and Wednesday, corresponding to, respectively, a smaller and larger width of the uncertainty range) would exceed 1 of 11 different indicated temperatures (5 above, 5 below, and 1 on the predicted temperature value for that day). All individual measurements were repeated four times, resulting in a total of 88 samples per participant: two x-positions or uncertainty widths (narrow and wide) × 11 different indicated temperatures × 4 repetitions. All 88 stimuli were presented in random order.

If $y_c(x)$ represents the vertical position of the center line at position x and $w(x)$ represents the width (i.e., the vertical distance between the outer edges of the uncertainty range) of the uncertainty interval at location x, the 11 (vertical) positions $y_i(x)$ of the sampling points are given by (see also Figure 6):

$$y_i(x) = y_c(x) + \frac{i}{6}.w(x)$$

for $i \in \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$. The width w increased monotonically (in this case linearly) with x along the center line.

To prevent stimulus familiarization, the shape of the center line was varied slightly across stimuli by randomly distributing the seven predicted

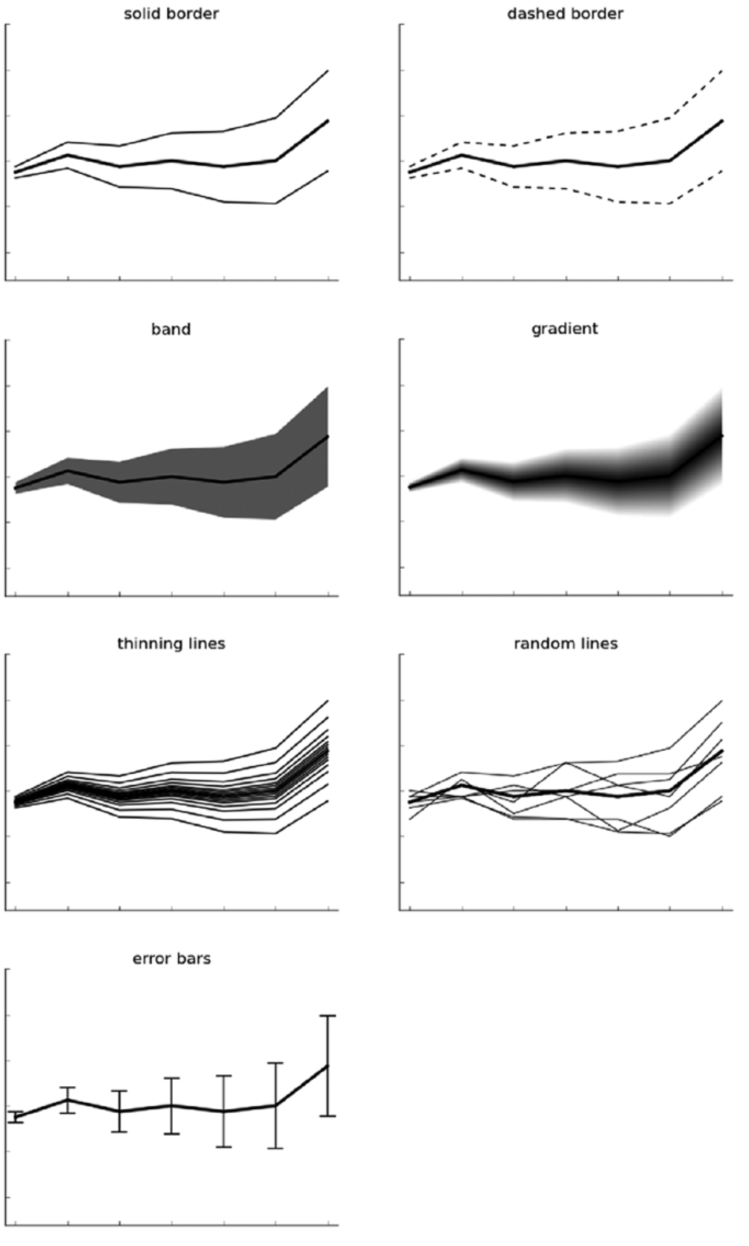

Figure 5. The seven graphical uncertainty visualization types used in the experiment. The thick black center line connects the seven data points (the 7-day predictions). Note that axis labels were present in the experiment (see Figure 4).

temperature values over the days of the week while keeping the width of the uncertainty interval fixed at each *x*-position (i.e., for any given day of the week, the *y*-value or uncertainty width was fixed, but the corresponding temperature value was randomly selected from the set of seven temperatures). This procedure yielded temperature curves with slightly varying shapes but similar and monotonously increasing uncertainty ranges (Figure 7).

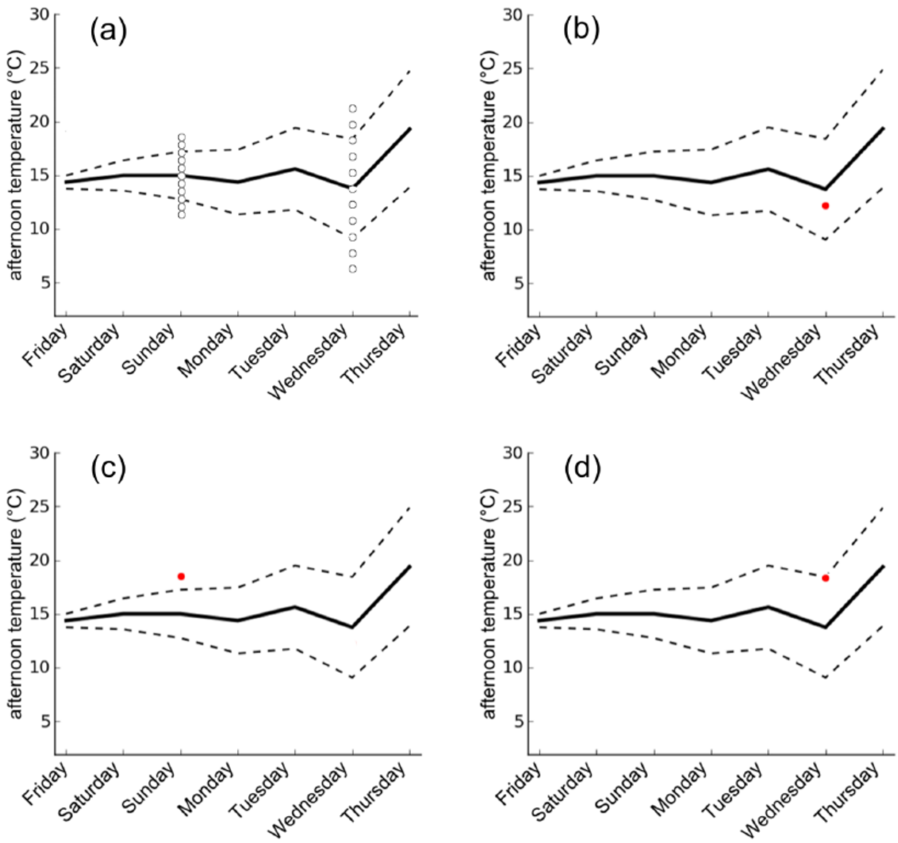

Figure 6. (a) The circles represent the 22 possible locations of the red dot. The dot was only presented at two positions along the horizontal (temporal) axis (respectively labeled *Sunday* and *Wednesday*). The width of the uncertainty interval at the location labeled *Wednesday* was about twice this width at the location labeled *Sunday*. At both horizontal positions (Sunday and Wednesday), the red dot could be presented at any of 11 vertical positions relative to the uncertainty range (5 below, 5 above, and 1 directly in the center of the uncertainty interval). (b-d) Three different instantiations of the stimulus.

Participants

In total, 140 people, naïve about the goal of the experiment, participated voluntarily (64 male, 76 female, age 18-57 years with a mean age of 38.1 years). The participants were recruited from the TNO database of volunteers and received a modest financial compensation for their participation. All participants gave their written consent prior to the experiment. Numeracy ranged from 1.5 to 5.9 (the Subjective Numeracy Scale ranges from 1 to 6; Fagerlin et al., 2007), with a mean score of 4.0. For a graphical overview of the education level (categorical, ranging from *primary education/ no education* to *PhD*) and numeracy of the participants, see Figure 8. Participants were randomly assigned to one of the seven visualization groups, such that each group consisted of 20 participants.

RESULTS AND DISCUSSION

Before analysis, we inspected the data for outliers. Most participants attributed probabilities such that probability was correlated with the (vertical) position of the sampling point (see Figure 9). However, some participants attributed the highest probability to sampling points on the center line and increasingly lower probabilities to sampling points located at larger distances from the center line (i.e., they adopted a “peaked” interpretation). We identified 16 participants that adopted such an interpretation.

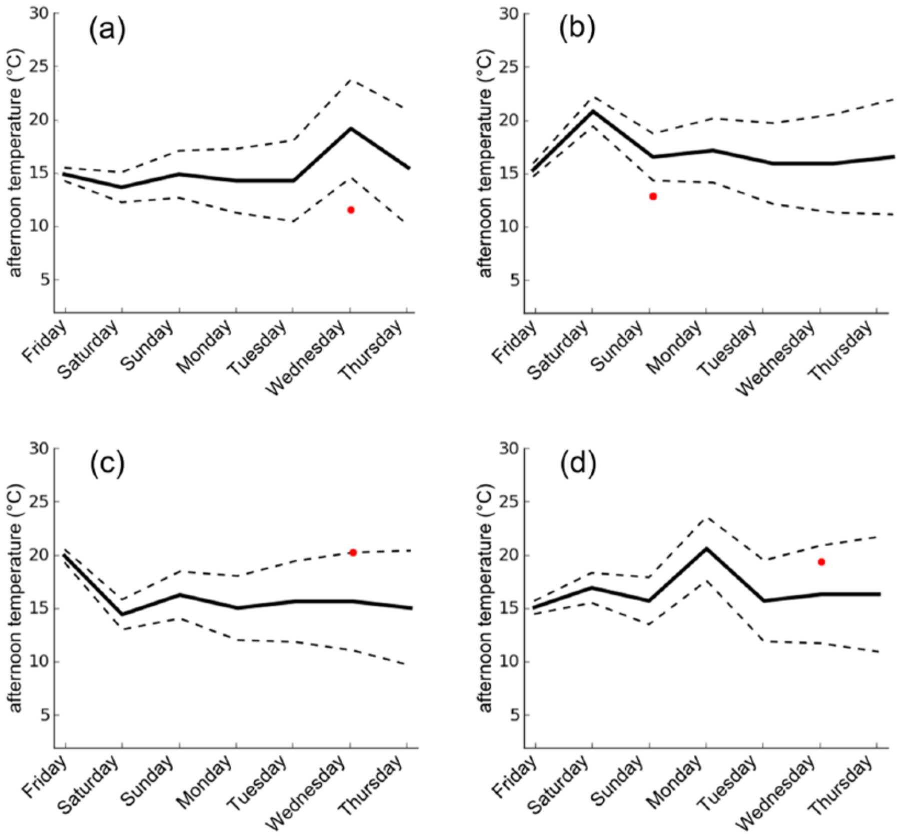

Figure 7. Four different instantiations of the stimulus. For each visualization type, the shape of the graphical temperature prediction varied throughout the experiment by randomly distributing the seven predicted temperature values over the days of the week, while keeping the width of the uncertainty interval fixed at each day.

There was no difference with regards to education level or numeracy between these and the other participants. Also, they were equally distributed across visualization types. Most likely, these participants inadvertently interpreted the range probability question as a point probability question. Also, there were eight participants who attributed probabilities such that probability was negatively correlated to the position of the sampling point; that is, they used a "colder than" rather than a "warmer than" interpretation. There was no difference with regards to education level or numeracy between these and the other participants. Also, they were equally distributed across visualization types.

The participants with a point probability or a "colder than" interpretation were regarded as erroneous and removed from the data set (see the Discussion section). Next, we examined if there were any extreme outliers among the remaining 116 participants. We examined the extreme outliers (data points that were more than 3 *SD*s above or below the mean) for each Position × Visualization Type combination. This process identified three additional participants as outliers. Results from these participants were also removed from the data set. All further analyses reported in this paper were performed on the data of the remaining 113 participants.

Distribution Shape

Similar to Tak et al. (2014) and to test our first hypothesis, we investigated the shape of the

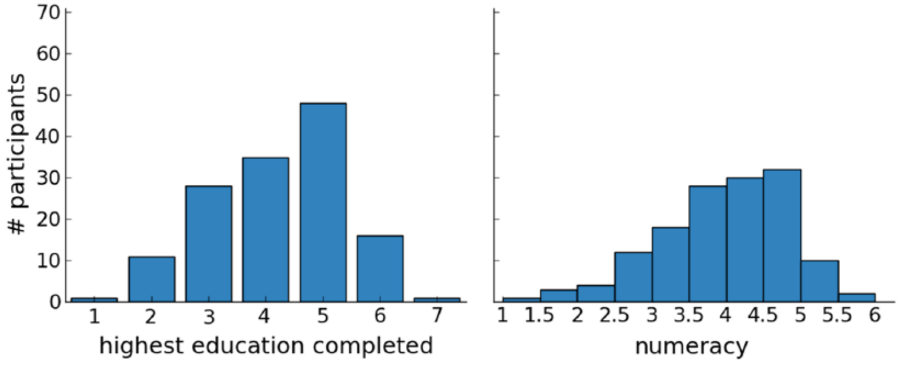

Figure 8. Highest education completed (1 = *primary education/no education*; 2 = *lower vocational education*; 3 = *lower secondary education*; 4 = *higher secondary education*; 5 = *BSc*; 6 = *MSc*; 7 = *PhD*) and numeracy of participants in the experiment. Bars show number of participants per category (education) or bin (numeracy).

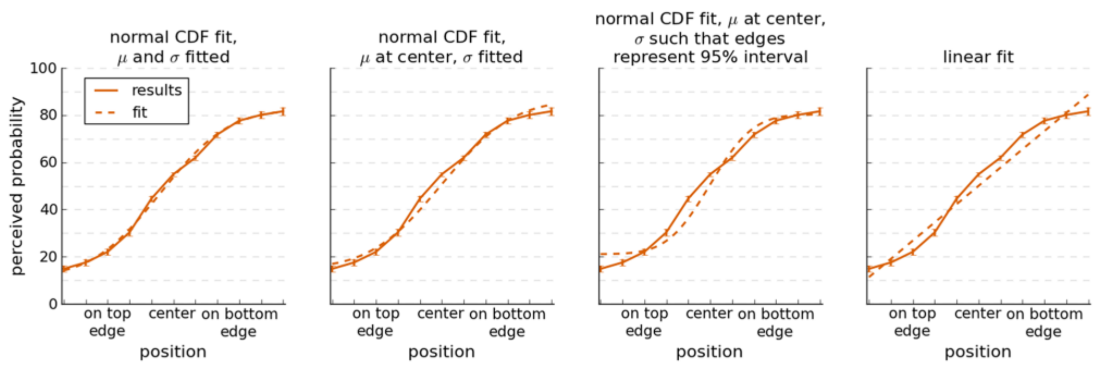

Figure 9. Results across visualization types and various fitted curves. Error bars denote ±1 SE.

perceived probability distribution. Participants may have a "linear" interpretation (consistent with a uniform distribution) or an interpretation that resembles a normal (cumulative) distribution (Tak et al., 2014). Knowledge of the shape of the inferred distribution function may help to explain effects like bias or misconceptions and may ultimately be deployed to produce visualizations that optimally convey the underlying distribution.

To be able to fit the data, we coded the position as 1-11, with 1 on the lowest position, 6 as the position on the center line, and 11 as the top position. Results of the fit in terms of μ and/or σ are given in "units" equaling one step size (see formula under methods) and with 6 units subtracted from the μ value to scale it from −5 to 5, with 0 at the center. We fitted the data (using SPSS 20.0) to the following:

- a normal cumulative distribution with its mean μ located at the center line and a width σ such that the 95% interval corresponds to the width of the uncertainty range (95% of the area under the normal distribution lies within 1.96 standard deviations of the mean, and the edges of the uncertainty range represent ±1.96 standard deviations of the mean),
- a normal cumulative distribution with its mean μ located at the center line and width σ fitted,
- a normal cumulative distribution with both mean μ and width σ fitted, and
- a linear (uniform) distribution, with μ located at the center line.

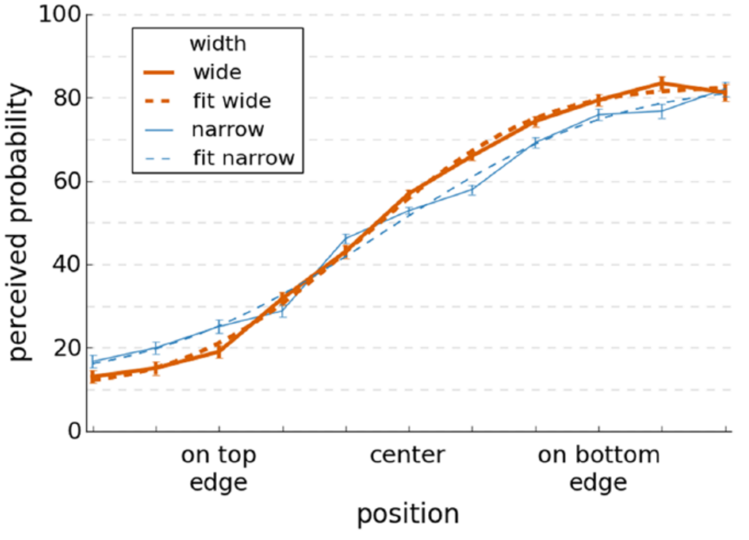

Figure 10. The significant Uncertainty Width × Position interaction. Error bars denote ±1 SE.

We find the following results (sorted by R^2 from high to low):

- normal CDF (cumulative distribution function) with μ and σ fitted: $R^2 = .761$ ($\mu = 0.6$ and $\sigma = 2.46$);
- normal CDF with σ fitted: adjusted $R^2 = .756$ ($\sigma = 2.57$);
- normal CDF with σ such that width of the uncertainty range represents a 95% interval: $R^2 = .743$; and
- linear (uniform) distribution: $R^2 = .739$.

Figure 9 shows these fits. This analysis shows that the results are best fitted by a normal curve with both μ and σ as free parameters. This confirms our first hypothesis (H1: People will apply a normal distribution to infer range probability from visualizations of temperature forecasts with uncertainty ranges when no further information is available, although note that given the nature of the distribution, a linear distribution will unavoidably have a relatively good fit as well. However, a priori, we did not expect a "bias" in the responses; that is, we expected the mean μ to be (close to) the center line because of the symmetrical nature of the stimuli. However, if we include μ as free parameter, the best fit is found for a value about 0.6 positions above the center line. This bias indicates that in general participants judge the probability of the temperature to exceed that at the center line to be above 50%. Similar to Tak et al. (2014), the fitted normal curve has a wider distribution than would have been expected if the outer lines corresponded with a 95% confidence interval: The participants' interpretation is (approximately) a 75% interval. However, as these results are an average of both the narrow and the wide uncertainty width conditions, it is informative to look at these two conditions separately as well to see if one of them is the main cause of the overdispersion (see Figure 10). In the narrow uncertainty width condition, we find that $\sigma = 2.87$; in the wide condition, $\sigma = 2.18$. According to a t test for independent samples, this difference is significant: $t(111) = 17.90$, $p < .001$. Though the results are even more dispersed in the narrow uncertainty width condition, both are notably more overdispersed than the results found by Tak et al. (2014), where $\sigma \approx 1.8$. Possibly, this is due to preconceptions people have about a "reasonable" uncertainty in temperature forecasts, in contrast to the study by Tak et al. (2014), where a case from an unfamiliar domain was used.

Visualization Type

We performed the fit with μ and σ as free parameters for each of the seven visualization types separately. Table 1 summarizes the results.

TABLE 1: Results of the Data Fit to a Normal Cumulative Function With μ and σ as Free Parameters for Each of the Visualization Types

Visualization Type	R^2	μ Mean (*SE*)	σ Mean (*SE*)
Band	.882	0.44 (.23)	2.70 (.38)
Dashed	.623	1.31 (.64)	2.74 (.88)
Error	.768	0.65 (.30)	2.42 (.45)
Gradient	.619	1.19 (.70)	2.98 (1.01)
Random	.809	0.24 (.21)	2.13 (.31)
Solid	.753	0.65 (.26)	2.09 (.37)
Thinning	.926	0.34 (.16)	2.37 (.24)

Differences between the visualization types were analyzed using a Tukey Honestly Significant Difference test. For μ, significant differences ($p < .05$) were found between dashed border (highest μ) and all other visualizations except gradient, and between gradient (second highest μ) and band, random lines, and thinning lines. For σ, significant differences ($p < .05$) were found for gradient (highest σ) and random lines and solid. These results support our second hypothesis (H2), that people's internal model depends on the visualization type. The mean values of μ indicate that all visualization types have a positive bias with the dashed border and gradient visualization as most "optimistic": With these visualizations, the occurrence of temperatures more than 1 unit above the center line is still perceived as 50-50. With regard to σ, the differences are small, with gradient having the most dispersed fit and random lines and solid having the steepest fit.

In addition and similar to Tak et al. (2014), we examined whether participants judged the probability of the "extreme values" (top-most and bottom-most points in Figure 6) as "impossible" (perceived probability 0) for the top-most point or as "absolutely certain" (perceived probability 100) for the bottom-most point. Overall, we find that 14% of responses on these outer points were either 0 or 100. This is a low score, but somewhat higher than the results found by Tak et al. (2014), who used 9 instead of 11 vertical locations, thereby putting the outmost sampling points at a smaller distance from the center line compared to the current study. The large majority of responses (86%) indicate that people nevertheless seem reluctant to consider any value to be "impossible" or "absolutely certain." There is a significant effect of visualization type on the occurrence of these extreme values, $\chi^2(6) = 14.3$, $p < .05$. From low to high, the visualization types scored as follows: error bars 5.9%, solid border 7.4%, dashed border 10.0%, band 14.1%, random lines 18.1%, thinning lines 19.6%, and gradient 23.4%, largely replicating the results for σ.

These results might suggest that the "density" of the fill within an uncertainty range determines how likely observers rate values outside the uncertainty range, with a denser fill leading to observers rating the extreme values as either "impossible" or "absolutely certain" more often. In other words, the denser the fill, the more likely it seems to observers that all values should fall within this area. This finding may be related to an earlier finding by Bisantz et al. (2009) that people assign regions that contrast most with the background to levels of meta-information that are most relevant to their task: If a task depended on the (un)certainty of information, participants assigned high contrast colors to highly (un)certain information. In our study, the uncertainty ranges with a relative dense fill (and thus a high contrast with the background) may have appeared more certain, thereby leading to observers rating the extreme values outside this area as either "impossible" or "absolutely certain" more often.

Summarizing, the data support our second hypothesis (H2), that people's internal model depends on the visualization type. Although we expected that the participants would estimate the probability that the temperature could exceed the center line to be 50% (i.e., $\mu = 0$), we found a consistent optimistic bias (participants consistently estimated this probability to be larger than 50%; i.e., $\mu > 0$) across the internal models for the different visualization types. The internal models for the different visualization types differ mainly in the size of this bias. Because the visual density of the pattern representing the area between the edges of the uncertainty range differed between the different uncertainty visualization types, we expected to find differences in the steepness of the (fitted) distribution curves. However, these differences were small and only significant for 3 of the 21 pairwise comparisons.

Uncertainty Width

Our third hypothesis (H3) was that the internal model would scale with the width of the uncertainty range; that is, we expected no effect of uncertainty width on the probabilities as a function of the relative position. However, our hypothesis H3 is not supported by the data, as we found significant differences between the wide and narrow width conditions on both μ and σ. For the narrow condition, $\mu = 0.40$ and $\sigma = 2.87$, and for the wide condition, $\mu = 0.73$ and $\sigma = 2.18$. According to a t test for independent samples, both differences are significant: μ, $t(111) = 13.87$, $p < .001$, and σ, $t(111) = 17.90$, $p < .001$ (see also Figure 10). The results show that the range probability estimates for points below the visualization center line are generally lower in the narrow uncertainty width condition than in the wide uncertainty width condition, whereas the reverse is true for the points above the visualization center line. This implies that for temperatures represented by points at the same relative distance from the center line, participants' judgments were affected by the uncertainty width. Note that this result is not trivial, as we intentionally chose the sampling positions to be at equal relative distances from the center line in both the narrow and wide uncertainty width conditions to avoid such an effect. This raises the question of whether participants' judgments are affected by the absolute distance of the points to the center line, as points at the same relative distance from the center line have a smaller absolute distance to this line in the narrow uncertainty width condition compared to the wide condition. However, Figure 11 shows that absolute distance does not account for the results found. Similar to the results by Sanyal et al. (2009), this shows that perceived uncertainty does not necessarily map linearly to visual features; a point x% outside a narrow uncertainty range is not perceived as (un)certain as a point x% outside a wide uncertainty range. Because there is no a priori reason to attribute different uncertainty estimates to temperatures represented by points at the same relative distance from the center line, this may indicate that either the interpretation of the uncertainty interval or the degree of trust in the forecast depends on the width of the uncertainty range. The current results do not allow us to distinguish between both explanations.

As for visualization type, we also looked at the occurrence of extreme values. The probability of the "extreme values" is affected by uncertainty width, $\chi^2(1) = 6.7$, $p < .05$, with 10% of the responses assigning either 0 or 100 to, respectively, the top-most and bottom-most points on the narrow uncertainty width, but 18% for the wide uncertainty width. Again, this effect can probably be attributed to the use of vertical locations relative to the uncertainty width, which puts the sampling points on the narrow uncertainty width at a lower absolute distance from the center line.

Numeracy

Our fourth hypothesis (H4) was that participant's numeracy relates to the shape of the internal model. To analyze the effect of numeracy, we divided the participants in a low numeracy group (numeracy score 3.5 and lower) and a high numeracy group (numeracy score above 3.5). The low numeracy group fit results in $\mu = 1.29$ and $\sigma = 2.75$ and the high numeracy group in $\mu = 0.44$ and $\sigma = 2.35$ (also see Figure 12). The differences in μ and σ are both significant, $t(111) = 16.38$, $p < .001$ and $t(111) = 5.64$, $p < .001$, respectively. The effect is similar to the finding of Tak et al. (2014) that participants with low numeracy adopt a "flatter" (more uniform) interpretation than those with high numeracy. As noted by Tak et al. (2014), numeracy may

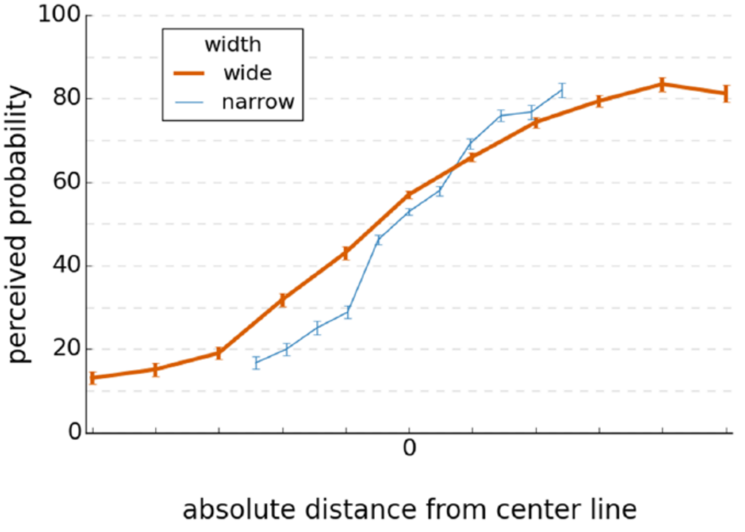

Figure 11. The significant Uncertainty Width × Position interaction by absolute distance from the center line. Error bars denote ±1 SE.

affect the mapping of values to the (nonlabeled) scale used in the experiment. Also, the visual interpretation of certainty may be affected by numeracy differences, with participants with relatively low numeracy having a "flatter," or more dispersed, interpretation than those with higher numeracy. In addition, the low numeracy group shows a larger (optimistic) bias than the high numeracy group. Hence, our hypothesis (H4) that the shape of the inferred distribution depends on people's level of numeracy is confirmed by the present data.

Comparison to Previous Work

Ibrekk and Morgan (1987) compared various displays to communicate cumulative distribution functions. The authors concluded that "[in] making judgments about probability intervals in displays that do not forcefully communicate a sense of probability density, people may use a linear proportion strategy which is equivalent to an assumption of uniform probability density" (pp. 527-528). Based on their results, the authors suggest that an error bar display (among others) may be an example of a display that does not properly communicate a sense of probability density. We can compare these results to the results of the current study, but we note that Ibrekk and Morgan (1987) used a 1D stimulus, whereas the current study uses a 2D display. Hence, the comparison between the current study and that of Ibrekk and Morgan (1987) may not be optimal. Fitting the results of the error bars visualization type in the current study to a normal cumulative distribution with its mean μ located at the center line and width σ fitted, and a linear (uniform) distribution through (μ, 50), with μ located at the center line, reveal a fit of R^2 = .762 to the normal distribution and R^2 = .744 to the linear distribution. Hence, we find no indication of a "linear proportion strategy" as found by Ibrekk and Morgan (1987). Possibly, error bar visualizations and/or probability density functions in general are more familiar to the general public nowadays than they were back in 1987, when Ibrekk and Morgan performed their studies.

An important feature of interval estimates is their alpha level—the probability or frequency with which the target value is expected to fall outside the interval. Alpha levels typically equal .10 or .05, respectively, corresponding to 90% or 95% certainty or reliability. Rinne and Mazzocco (2013) studied the effects of alpha level and numeracy on the distribution of inferences that people make when given textual descriptions of uncertainty intervals. In their experiment, participants received prediction intervals

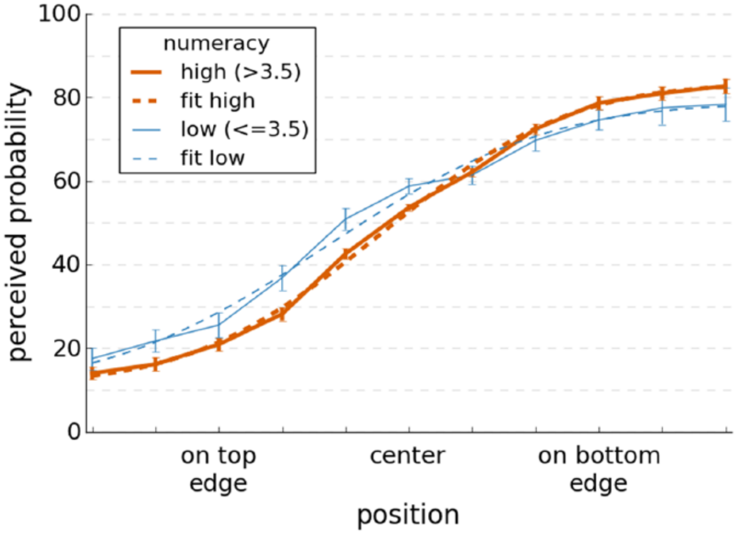

Figure 12. The significant Position × Numeracy interaction. Error bars denote ±1 SE.

for fictitious towns' annual rainfall totals (assuming approximately normal distributions) and estimated probabilities that future totals would be captured within varying margins about the mean, indicating the approximate shapes of their inferred probability distributions. The authors find that low alpha levels in particular led to overdispersed interpretations (compared to moderate alpha levels like .25, which led to the most accurate distributional inferences). Future work combining the current study and the work by Rinne and Mazzocco (2013) could shed more light on this question. Rinne and Mazzocco (2013) also examined the effect of numeracy. They found that highly numerate participants are more likely to infer (approximately) uniform distributions. In the current study, we find no evidence of this. For participants with both high and low numeracy, we find a better fit for a normal cumulative distribution than a linear (uniform) distribution, though (not unexpected given the nature of the distribution) the difference is not very large.

CONCLUSION AND FUTURE WORK

The results of this study confirm earlier findings (Tak et al., 2014) that, in the absence of any (textual) explanation of uncertainty range, people will apply an internal model of the uncertainty distribution that closely resembles a normal (cumulative) distribution. In contrast to previous work, we find only a marginal effect of visualization type on the disparity of the perceived probability, but the perceived probability of "extreme values" (i.e., those far outside the uncertainty range) is affected by the visualization type, with denser fills leading to higher perceived probability of values within that area. In addition, we do find large differences in the bias, indicating a systematic shift of the perceived probability toward an optimistic view, which is larger than 1 unit for the dashed lines and gradient visualizations.

Perceived probability also depends on the width of the uncertainty range: The probability of values with equal relative difference from the center value of the probability range is judged differently for wide and for narrow uncertainty ranges. This means that observers take both the relative and absolute distance to the center line into account and the internal model does not scale with uncertainty width. Finally, as also noted by Tak et al. (2014), the internal model of the uncertainty distribution relates to a participant's numeracy, though the cause of this effect is still unclear.

We note that compared to Tak et al. (2014) the current study had a relatively high number of outliers. We identify two possible reasons for

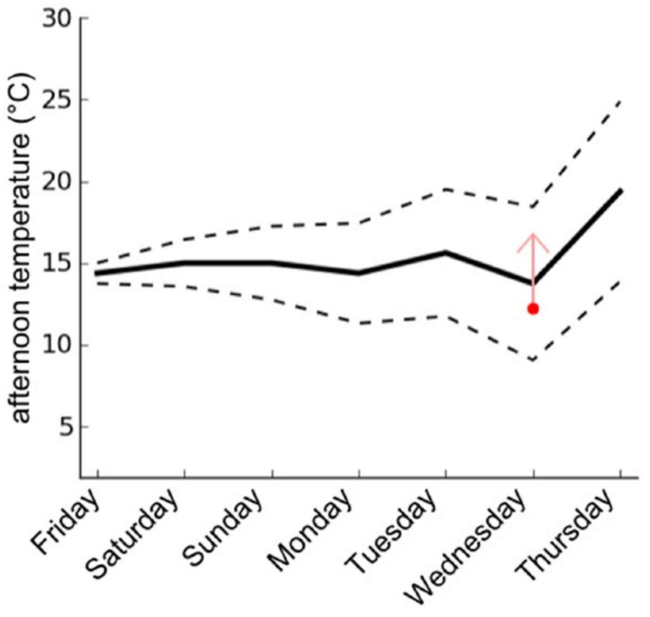

Figure 13. Proposed stimulus for future research.

this effect. First, the "range probability" question used in the current study may be more difficult to answer than the "point probability" question used by Tak et al. (2014). In the case of the current study, the range probability question led to counterintuitive results in the sense that points with low (high) temperature values have relatively high (low) perceived probability. Given that the question was also one with a positive connotation ("warmer than"), this may have confused participants (this is supported by the fact that eight participants seemed to have adopted a "colder than" interpretation). Second, some confusion may also have stemmed from the fact that the stimuli used a red dot to indicate the temperature, which may have suggested a point probability question (this is supported by the fact that 16 participants seemed to have adopted a point probability interpretation). Possibly, the stimuli should have used a visual method more appropriate for a range question, such as an upward pointing arrow (see Figure 13). This could also address the issue of participants incorrectly interpreting the question as a "colder than" question.

The results are consistent with previous work (Tak et al., 2014) in the sense that once again the results show that the internal model of the uncertainty distribution closely resembles a normal distribution. Therefore, it seems less relevant whether a point or a (possibly more difficult) range probability question is used. Also, the internal model does not appear to be strongly affected by familiarity with the subject of the visualization in the sense that the results are still best described by a normal curve. However, the results in the current study are notably more overdispersed than the results found by Tak et al. (2014). Possibly, this stems from the fact that a case from a more familiar domain was used in the current study.

We also identify some implications for makers or designers of uncertainty visualizations. Our results imply that low alpha levels may not be the best choice for uncertainty visualizations, as observers' interpretations are more dispersed. It is doubtful that textual labeling of the uncertainty interval will resolve this problem (Rinne & Mazzocco, 2013). We also found a systematic bias in our data, indicating an optimistic interpretation of the weather forecast. This bias has not been reported before and general recommendations can only be made after the effect has been replicated and found in other domains than weather forecasts. In terms of the different visualization types, it is unclear what the "best" choice is. However, the effects of the density of the fill on the interpretation of extreme values should be taken into account when choosing a particular visualization, especially when these unlikely events are of the type "low probability, high impact" (Bussiere & Fratzscher, 2008). Finally, the different results on the narrow and the wide uncertainty width again (Sanyal et al., 2009) show that perceived uncertainty does not necessarily map linearly to visual features and that testing of the interpretation of uncertainty visualizations prior to dissemination is important, as the intentions of the designer do not necessarily match the interpretation of the viewer.

In this study, we operationalized range probability as the probability that the temperature exceeds a given threshold value, thus focusing on the range between this threshold and infinity, which is common in temperature communication. Because probabilistic forecasts are typically used for planning and risk management where different ranges may require different actions, future studies should also investigate how observers estimate the probability that a given entity (e.g., temperature, precipitation or water levels, and wind force) is within a given restricted range. Such an approach may also resolve the question of whether the bias towards an optimistic interpretation results from the instruction given to the participants ("estimate the probability that the temperature will exceed a given threshold value"), which may have focused their attention on higher temperatures.

Future work could also focus on the effect of (different types of) labeling and/or textual explanation of the uncertainty range. To isolate the effect of visual representation, the current study did not include (textual) explanation of the uncertainty range (e.g., "95% confidence interval"). The effect of labeling or even what is the most appropriate labeling (particularly for nonexperts or people with low numeracy) remains unknown. As suggested before (Savelli & Joslyn, 2013), labeling may help to prevent the automatic deterministic interpretation of predictive forecast intervals.

People have an unconscious tendency to use previously considered standards as a reference when making numeric estimates (an effect known as anchoring; Tversky & Kahneman, 1974). Both the center line and the red reference dot in the stimuli used in this study may have induced anchoring effects. Although previous research has shown that the effect of anchoring on numerical estimates from probabilistic weather forecasts is reduced by providing both the upper and lower bounds of the predictive uncertainty ranges (Joslyn et al., 2011), there may still have been a net effect of the center line and red dot in this study. Therefore, it would be interesting to investigate a condition without a center line and one in which the red dot is absent or replaced by a less salient marker.

Also, it would be interesting to investigate if providing uncertainty information in a verbal format instead of a graphical representation results in similar internal models and could serve to reduce the optimistic estimation bias observed in this study. A condition without uncertainty information could serve to test if participants also assume a roughly normal distribution when only a central line is available, based on previous experiences with forecasts and their outcomes. Another suggestion for future work is the inclusion of experts, as both Tak et al. (2014) and the current study used nonexpert participants. Previous work (Belia et al., 2005) suggests that the results for experts may be different from those for nonexperts. Finally, the relation between the internal model that observers appear to apply and numeracy seems consistent but remains unexplained. Future work could examine this finding further.

Our study shows that the width and density of graphical representations of uncertainty ranges affect range probability estimates and do so differently for estimates relative to reference values in, respectively, the upper or lower part of an uncertainty range. Though further study is needed to establish the generalizability of this result, we suspect that the effects found here are likely to hold for a wider range of visualizations. This may have practical implications for graphical forecasts used in different areas like agriculture, flood management, health care, finance, and many other decision-making contexts where incorrect inferences from range estimates may lead to suboptimal decisions. Further research can provide knowledge on the nature of the effects found here, which may in turn lead to more effective presentations of uncertainty ranges to diverse populations in a variety of judgment and decision-making contexts.

REFERENCES

Andre, A. D., & Cutler, H. A. (1998). Displaying uncertainty in advanced navigation systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *41*(2), 31-35.

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

Bisantz, A. M., Stone, R. T., Pfautz, J., Fouse, A., Farry, M., & Roth, E., et al. (2009). Visual representations of meta-information. *Journal of Cognitive Engineering and Decision Making*, *3*(1), 67-91.

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.

Brolese, A., & Huf, S. (2006). Visualizing uncertainty to improve operators' spatial proximity judgments in uncertain surroundings. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(3), 294-298.

Bussiere, M., & Fratzscher, M. (2008). Low probability, high impact: Policy making and extreme events. *Journal of Policy Modeling*, *30*(1), 111-121.

Cedilnik, A., & Rheingans, P. (2000). Procedural annotation of uncertain information. In *Proceedings of visualization 2000* (pp. 77-84). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Craig, G. C. (2012). Probabilistic weather forecasting. In U. Schumann (Ed.), *Atmospheric physics* (pp. 661-673). Berlin-Heidelberg, Germany: Springer.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics*, *29*(3), 89-93.

de Elía, R., & Laprise, R. (2005). Diversity in interpretations of probability: Implications for weather forecasting. *Monthly Weather Review*, *133*(5), 1129-1143.

Demeritt, D., Cloke, H., Pappenberger, F., Thielen, J., Bartholmes, J., & Ramos, M. H. (2007). Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards*, *7*(2), 115-127.

Demeritt, D., Nobert, S., Cloke, H. L., & Pappenberger, F. (2013). The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrological Processes*, *27*(1), 147-157.

Demuth, J., Morss, R. E., Lazo, J. K., & Stewart, A. E. (2007). Assessing how the U.S. public understands and uses weather forecast uncertainty information. In *Weather and forecasting: Vol. 3.3. Proceedings of the 16th Conference on Applied Climatology* (pp. 1-4). San Antonio, TX: American Meteorological Society.

Ehlschlaeger, C. R., Shortridge, A. M., & Goodchild, M. F. (1997). Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, *23*(4), 387-395.

Ehrendorfer, M. (1997). Predicting the uncertainty of numerical weather forecasts: A review. *Meteorolgische Zeitschrift, Neue Folge*, *6*, 147-183.

Evans, B. J. (1997). Dynamic display of spatial data-reliability: Does it benefit the map user? *Computers & Geosciences*, *23*(4), 409-422.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making*, *27*(5), 672-680.

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

Garcia-Retamero, R., & Galesic, M. (2010). Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science & Medicine*, *70*(7), 1019-1025.

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, *83*(0), 27-33.

Gershon, N. (1998). Visualization of an imperfect world. *IEEE Transdactions on Computer Graphics and Applications*, *18*(4), 43-45.

Gigerenzer, G., Hertwig, R., Van Den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis*, *25*(3), 623-629.

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*(5746), 248-249.

Griethe, H., & Schumann, H. (2005). Visualizing uncertainty for improved decision making. In *Proceedings of the 4th International Conference on Perspectives in Business Informatics Research (BIR 2005)* (pp. 1-11). Skövde, Sweden: University of Skövde.

Hildon, Z., Allwood, D., & Black, N. (2012). Impact of format and content of visual display of data on comprehension, choice and preference: A systematic review. *International Journal for Quality in Health Care*, *24*(1), 55-64.

Ibrekk, H., & Morgan, M. G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, *7*(4), 519-529.

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126-140.

Joslyn, S., & LeClerc, J. (2013). Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, *22*(4), 308-315.

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*(2), 133-147.

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, *17*(2), 180-195.

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*(4), 342-353.

Juanchich, M., Teigen, K. H., & Gourdon, A. l. (2013). Top scores are possible, bottom scores are certain (and middle scores are not worth mentioning): A pragmatic view of verbal probabilities. *Judgment and Decision Making*, *8*(3), 345-364.

Kirlik, A. (2007). Lessons learned from the design of the decision support system used in the Hurricane Katrina evacuation decision. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 253-257.

Kootval, H. (2008). *Guidelines on communicating forecast uncertainty* (technical report WMO/TD No. 4122). Geneva, Switzerland: World Meteorological Organization.

Leith, C. E. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, *102*(6), 409-418.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013). The heuristic interpretation of box plots. *Learning and Instruction*, *26*, 22-35.

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

Marimo, P., Kaplan, T. R., Mylne, K., & Sharpe, M. (2012). *Communication of uncertainty in weather forecasts (MPRA Paper No. 38287)*. Exeter, UK: University of Exeter Business School.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Weather and Forecasting*, *23*(5), 974-991.

Morss, R. E., Lazo, J. K., & Demuth, J. L. (2010). Examining the use of weather forecasts in decision scenarios: Results from a US survey with implications for uncertainty communication. *Meteorological Applications*, *17*(2), 149-162.

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.

Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, *3*(3), 209-227.

Nadav-Greenberg, L., Joslyn, S. L., & Taing, M. U. (2008). The effect of uncertainty visualizations on decision making in weather forecasting. *Journal of Cognitive Engineering and Decision Making*, *2*(1), 24-47.

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

Pappenberger, F., Stephens, E., Thielen, J., Salamon, P., Demeritt, D., & van Andel, S. J., et al. (2013). Visualizing probabilistic flood forecast information: Expert preferences and perceptions of best practice in uncertainty communication. *Hydrological Processes*, *27*(1), 132-146.

Peachy, J. A., Schultz, D. M., Morrs, R., Roebber, P. J., & Wood, R. (2013). How forecasts expressing uncertainty are perceived by UK students. *Weather*, *68*(7), 176-181.

Pöthkow, K., & Hege, H. C. (2011). Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*, *17*(10), 1393-1406.

Potter, K., Rosen, P., & Johnson, C. R. (2012). From quantification to visualization: A taxonomy of uncertainty visualization approaches. In A. M. Dienstfrey & R. F. Boisvert (Eds.), *IFIP Advances in information and communication technology Volume 377. Uncertainty quantification in scientific computing* (pp. 226-249). Berlin-Heidelberg, Germany: Springer.

Ramos, M. H., van Andel, S. J., & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, *17*(6), 2219-2232.

Richardson, D. S., Bidlot, J., Ferranti, L., Ghelli, A., Haiden, T., & Hewson, T., et al. (2011). *Verification statistics and evaluations of ECMWF forecasts in 2010-2011* (ECMWF technical memorandum, Vol. 564). Reading, UK: European Centre for Medium-Range Weather Forecasts.

Rinne, L. F., & Mazzocco, M. M. M. (2013). Inferring uncertainty from interval estimates: Effects of alpha level and numeracy. *Judgment and Decision Making*, *8*(3), 330-344.

Riveiro, M. (2007). Evaluation of uncertainty visualization techniques for information fusion. In *Proceedings of the 10th International Conference on Information Fusion (ICIF'07)* (pp. 1-8). Québec, Canada: International Society for Information Fusion.

Roulston, M. S., Bolton, G. E., Kleit, A. N., & Sears-Collins, A. L. (2006). A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting*, *21*(1), 116-122.

Roulston, M. S., & Kaplan, T. R. (2009). A laboratory-based study of understanding of uncertainty in 5-day site-specific temperature forecasts. *Meteorological Applications*, *16*(2), 237-244.

Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., & Moorhead, R. (2009). A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1209-1218.

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*(6), 1421-1430.

Savelli, S., & Joslyn, S. L. (2009). Visualizing temperature forecast uncertainty for a non-expert web audience. In *Proceedings of the 9th International Conference on Naturalistic Decision Making NDM9* (pp. 283-284). London: British Computer Society.

Savelli, S., & Joslyn, S. (2013). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, *27*(4), 527-541.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393-1400.

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.

Tak, S., Toet, A., & van Erp, J. B. F. (2014). The perception of visual uncertainty representation by non-experts. *IEEE Transactions on Visualization and Computer Graphics*, *20*(6), 935-943.

Teigen, K. H., & Filkuková, P. (2013). Can>will: Predictions of what can happen are extreme, but believed to be probable. *Journal of Behavioral Decision Making*, *26*(1), 68-78.

Teigen, K. H., Juanchich, M., & Filkuková, P. (2013). Verbal probabilities: An alternative approach. *The Quarterly Journal of Experimental Psychology*, *67*(1), 124-146.

Teigen, K. H., Juanchich, M., & Riege, A. H. (2013b). Improbable outcomes: Infrequent or extraordinary? *Cognition*, *127*(1), 119-139.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Whitaker, R. T., Mirzargar, M., & Kirby, R. M. (2013). Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2713-2722.

Zuk, T., & Carpendale, S. (2007). Visualization of uncertainty and reasoning. In A. Butz, B. Fisher, A. Krüger, P. Olivier, & S. Owada (Eds.), *Vol. lecture notes in computer science Volume 4569/2007. Smart Graphics* (pp. 164-177). Berlin-Heidelberg, Germany: Springer.

Susanne Tak received her PhD in human–computer interaction from the Department of Computer Science and Software Engineering of the University of Canterbury, Christchurch, New Zealand, in December 2011. She currently works as an education developer, at the Freudenthal Institute for Science and Mathematics Education, Utrecht University, the Netherlands. Her research interests include information visualization, cognitive psychology, human–computer interaction, and human problem solving.

Alexander Toet is a senior scientist in the Department of Perceptual and Cognitive Systems at The Netherlands Organization for Applied Scientific Research TNO. He conducts research in areas of multisensory and affective perception, visual saliency, clutter, and multimodal image fusion. He is a fellow of the SPIE and Area Editor of the *ISIF Journal of Advances in Information Fusion*. He received his PhD in physics from the University of Utrecht in the Netherlands.

Jan van Erp is a senior scientist and program manager in the Department of Perceptual and Cognitive Systems at The Netherlands Organization for Applied Scientific Research TNO and is scientific director of the BrainGain consortium. His research focuses on advanced human–computer interaction, including multimodal interaction, augmented reality, and brain–computer interfaces. He received his PhD in computer science from Utrecht University.