Examining Interpretation Strategies for Multiple Forecast Visualizations with Two and Four Forecasts

Lace Padilla
l.padilla@northeastern.edu
Northeastern University
Boston, USA

Racquel Fygenson
fygenson.r@northeastern.edu
Northeastern University
Boston, USA

Connor Wilson
wilson.conn@northeastern.edu
Northeastern University
Boston, USA

Kristi Potter
Kristi.Potter@nrel.gov
National Renewable Energy Laboratory
Denver, USA

Spencer C. Castro
scastro39@ucmerced.edu
University of California, Merced
Merced, USA

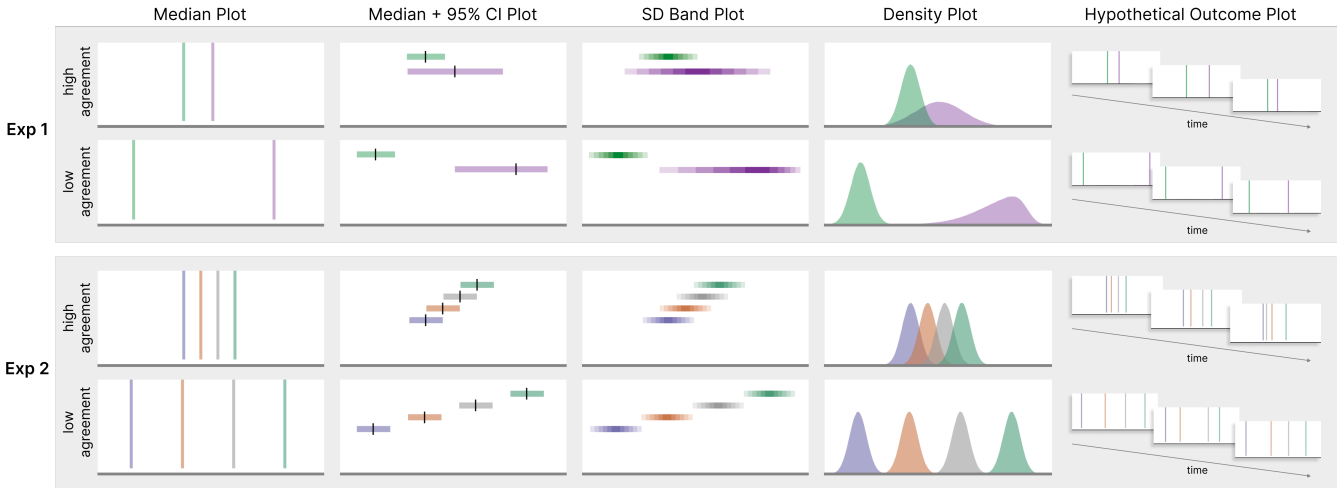


Figure 1: Visualization techniques evaluated in Experiments 1 and 2. Columns show the five visualization types: median plots, 95% confidence intervals, standard deviation (SD) bands, density plots, and hypothetical outcome plots (HOPs). Rows display forecast agreement conditions (high vs. low). Experiment 1 (top) includes two forecast distributions, and Experiment 2 (bottom) extends to four forecast distributions under both agreement levels.

ABSTRACT

Multiple forecast visualizations (MFVs) present curated sets of forecasts to support decision-making under uncertainty. However, the research community knows little about how people interpret and integrate competing forecasts. In this study, we investigate the strategies individuals use when predicting hypothetical future events with MFVs across five visualization types (median, 95% CIs, standard deviation intervals, density plots, and hypothetical outcome plots) and multiple probability distributions in two preregistered experiments (n = 500 each). Analysis of 18 participant strategies and open responses shows that whereas many participants attempted to visually average across forecasts, others adopted a winner-takes-all approach (*e.g.*, selecting a single forecast as the most likely outcome), which deviates from rational agent expectations. We also observed reliance on visual artifacts, such as intersection points or end caps. These findings underscore the complexity of interpreting a range of forecasts and help explain why individuals may privilege particular predictions in real-world decision contexts.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in visualization**; **Empirical studies in visualization**; **Visualization techniques**.

KEYWORDS

Multiple Forecast Visualization, uncertainty, distributions, density plots, hypothetical outcome plots

ACM Reference Format:
Lace Padilla, Racquel Fygenson, Connor Wilson, Kristi Potter, and Spencer C. Castro. 2018. Examining Interpretation Strategies for Multiple Forecast Visualizations with Two and Four Forecasts. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 19 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Over the past 50 years, advances in data availability, computing power, and statistical methodology have led to a dramatic expansion in the number of forecasting models available across scientific and applied domains [51]. As a result, analysts, policymakers, and the general public routinely encounter multiple forecasts for the same event. This rising prevalence has intensified the need for effective methods of communicating collections of forecasts in ways that align with a communicator's goals. Emerging work on *multiple forecast visualizations* offers a promising approach, enabling users to view an array of independent forecasts within a single display, often outperforming common summary plots such as 95% confidence intervals [40, 41]. However, progress on multiple forecast visualizations is hindered by our incomplete understanding of the strategies people use when interpreting multiple forecasts simultaneously. Absent this foundational knowledge, designers cannot reliably create multiple forecast visualizations that support interpretations consistent with their intended communication goals.

Furthermore, although mathematical methods for combining forecasts are well-established [9], scholars emphasize that no single method is universally optimal because appropriateness of approaches depends on factors such as data structure, model quality, and the specific forecasting task, among other considerations [51]. Given these contingencies, it is crucial to understand how visualization techniques afford particular interpretation strategies so that forecasters and designers can select formats that best support their objectives. To address this gap, we present a context-free sequence of experiments to identify the strategies the general public uses when interpreting multiple forecasts across common visualization techniques. By deliberately removing domain context, we establish a baseline set of perceptual and cognitive tendencies that designers can use to anticipate how their forecasting contexts may interact with, amplify, or attenuate these baseline strategy patterns.

To investigate how people interpret multiple forecast visualizations, we conducted two preregistered experiments using five common visualization types: median plots, 95% confidence intervals, standard deviation bands, density plots, and hypothetical outcome plots (HOPs) (shown in Figure 1). Experiment 1 examined how participants interpreted two forecasts, including how their strategies shifted when forecasts were skewed or showed substantial disagreement. Experiment 2 tested whether these patterns persisted when participants evaluated four forecasts. Together, the experiments provide converging evidence that people rely on a diverse set of strategies, and that these strategies shift systematically with both the visualization technique and the properties of the forecasts. All data, analyses [osf.io/rwnt8], and preregistrations [Exp 1: osf.io/pfjwe, and Exp 2: osf.io/3mcjp] are available on the Open Science Framework (OSF). The major contributions include:

- Converging empirical evidence from a rigorous controlled study that participants' strategies for evaluating multiple forecast visualizations are influenced by visualization type and forecast properties.
- Providing empirical evidence that a significant portion of lay audiences adopt a winner-takes-all approach when interpreting visualized forecasts, favoring a single forecast as the most likely outcome instead of combining forecasts.
- Documenting instances where individuals switch strategies in predictable ways based on forecast properties.
- Providing evidence that, under certain conditions, participants rely on visual artifacts (*e.g.,* intersection points, end caps) while neglecting distributional information, introducing unintended variability in judgments.

2 RELATED WORK

In uncertainty visualization, designers typically choose between two main approaches [42]. One approach uses summary statistics, such as confidence intervals or mean plots, to present information succinctly. Despite their widespread adoption (*e.g.,* confidence intervals appeared in 60% of COVID-19-related line charts [57]), they are often misunderstood [3, 8, 11, 18, 19, 24–26, 44, 46, 47], even after educational interventions [4] and regardless of viewers' expertise [3]. A key issue, identified as *deterministic construal error*, occurs when viewers misinterpret uncertainty as deterministic rather than probabilistic [25]. For example, in hurricane forecast visualizations, viewers often misinterpret a confidence interval as an impact zone, rather than as distributional information [43, 47].

Alternatively, designers can use distributional visualizations, such as gradient, violin [11], quantile dot [30], and ensemble plots [34], which have been shown to enhance comprehension and performance compared to summary statistics [8, 11, 14, 21, 26, 27, 30, 44, 47], and textual descriptions of distributions and visualizations without uncertainty [8, 14]. These distributional visualizations can be categorized as encoding uncertainty *explicitly* or *implicitly* [42]. Explicit encodings, such as quantile dot [30] and density plots, utilize graphical elements to represent confidence levels or probabilities directly. In contrast, implicit encodings convey uncertainty through the representation of various potential outcomes without direct quantification [12], as seen in ensemble charts [34], hypothetical outcome plots (HOPs [22]), and multiple forecast visualizations [41].

Multiple forecast visualizations offer a lesser-studied method for communicating implicit uncertainty by showing variance or agreement among multiple forecasts. This approach conveys the diversity, form, and density of future projections [41]. Unlike ensemble or hypothetical outcome plots that generate predictions from a statistical model [22, 32–34], **multiple forecast visualizations display forecasts from different independent sources or forecasting groups**, enabling direct comparison of various predictive models in a single visualization. Because multiple forecast visualizations display multiple independent forecasts, their design involves both visualization and data collation choices, making them a case where data selection and information visualization intersect.

Decision-making strategies vary widely when people interpret multiple forecast visualizations, shifting with the number, distribution, and context of the forecasts. For instance, work on COVID-19

mortality forecasts found that viewer trust increases as the number of forecasts grows (up to roughly six to nine) before stabilizing, and that worst-case outliers can bias judgments unless embedded within a larger set of forecasts [41]. Sarma, Hedayati, and Kay [48] similarly observed that when viewing cumulative distribution plots, participants sometimes overweighted a uniformly distributed worst-case forecast, whereas in other situations they discounted a lone outlier when most forecasts clustered near one bound. McKenzie et al. [35] identified additional simple heuristics used when comparing competing predictions and highlighted the need to study such divergent strategies more systematically.

The present study addresses this gap by evaluating a broad set of interpretation strategies across a wider range of visualization types in a controlled context free setting. Prior work has typically focused on specific application domains (for example GPS navigation [35], disaster risk management [48], or pandemic forecasting [40, 41]), which limits generalization and provides little insight into how visualization design alone shapes strategy use. In contrast, our approach establishes a general baseline for how interpretation strategies vary across common visualization techniques. We tested the full strategy set described in subsection 2.1 using five visualization types shown in Figure 1. Because each technique introduces distinct visual features that may cue different strategies, examining only a small number of forms can miss important behaviors. Indeed, previous studies often compared only two techniques at a time, such as ensemble plots versus probability boxes [48], confidence intervals versus Gaussian fades [35], or median lines alongside confidence intervals [40, 41]. By testing a larger set of visualization types, evaluating an exhaustive strategy list, and removing strong contextual cues, we aim to advance theoretical understanding of the strategies people use when comparing multiple forecasts and the visual features that support them.

2.1 Strategies and Perceptual Proxies

In the field of forecast combination, there is ongoing debate about the "correct" statistical method for aggregating forecasts, with no universally accepted approach because factors such as context and forecaster characteristics substantially influence the choice of technique (*e.g.,* [51, 56]). Given these methodological debates, we do not assume a single correct method for combining distributions. Instead, we focus on identifying the strategies participants use to interpret multiple forecast visualizations and the forecast properties that shape those strategies.

The absence of a universal method for combining forecasts motivates our goal of examining how people naturalistically interpret visual representations of multiple forecasts. Prior visualization research has examined how viewers interpret data visualizations, notably through the examination of *perceptual proxies* [23, 39, 55]. Perceptual proxies, or visualization features used to solve tasks [23], have been identified in bar charts for assessing maximums, means, ranges, correlations, and differences [23], and in scatter plots [39, 55]. Identifying visualization features that guide viewers' interpretations enables the comparison and contrast of candidate visual proxies to determine which one best influenced participants' judgments. Extending this foundation, related work has documented analogous strategy use in the context of uncertainty visualizations (*e.g.,* [26, 36]). This research not only identifies reasoning strategies but also highlights how participants rely on specific visual features, such as the mean in confidence intervals or density plots, or use visual distance between results as a proxy for effect size [26]. More studies find that participants' judgments can be influenced by features of uncertainty representation, such as the end caps of error bars [24].

We investigate participants' visual strategies for comparing multiple forecast visualizations. Using the perceptual proxies approach [23, 39, 55], we define 18 candidate interpretation strategies, which we organize into three groups, shown in Figure 2 (**Data-Driven Mental Averaging**, **Winner-Takes-All**, and **Heuristic Visual Artifact**). In the following passages, we detail each strategy and the rationale behind it.

Data-Driven Mental Averaging Strategies*: A strategy in which viewers mentally combine forecast distributions by internally averaging key statistical features.* Prior work with numerical stimuli shows that people often average pairs of numbers [6], to the point that scholars describe it as "a rule that is (almost) universally invoked by human judges" [6]. Visualization research similarly finds that participants frequently average forecasts and sometimes assign more weight to those with smaller standard deviations [17]. Mental averaging is therefore one of the most consistently observed strategies across numerical and visual contexts. Because it can be implemented in multiple ways, we enumerate all plausible variants to assess which forms participants use for each visualization type (shown in Figure 2, left column).

The simplest forms of mental averaging combine the central tendencies of the forecasts, such as taking the mean, median, or mode of the two distributions (Figure 2, left column, top rows). More sophisticated approaches integrate additional distributional properties, including standard deviation or skew, producing a weighted mental average that reflects differences in shape. A well-established formulation of such a weighted combination is the linear opinion pool (LOP; Equation 1) [2, 50], following the adaptation from [9].

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta) \tag{1}$$

Here, each expert i's belief about the uncertain parameter θ is represented by a probability distribution $p_i(\theta)$, and $p(\theta)$ denotes the aggregated distribution reflecting a consensus across all n forecasts. The weighting factor w_i for each forecast distribution is defined as the reciprocal of the variance of $p_i(\theta)$ and is normalized so that the w_i sum to unity, following the method of Greis et al. [17]. Throughout, we use p as a general notation that can denote either a probability mass function (for discrete variables) or a probability density function (for continuous variables).

Below are the four Data-Driven Mental Averaging strategies we tested, which are shown in the left column of Figure 2:

(1) Mentally averaging the means of the two distributions.
(2) Mentally averaging the medians of the two distributions.
(3) Mentally averaging the modes of the two distributions.
(4) Mentally computing a weighted average using the linear opinion pool formulation.

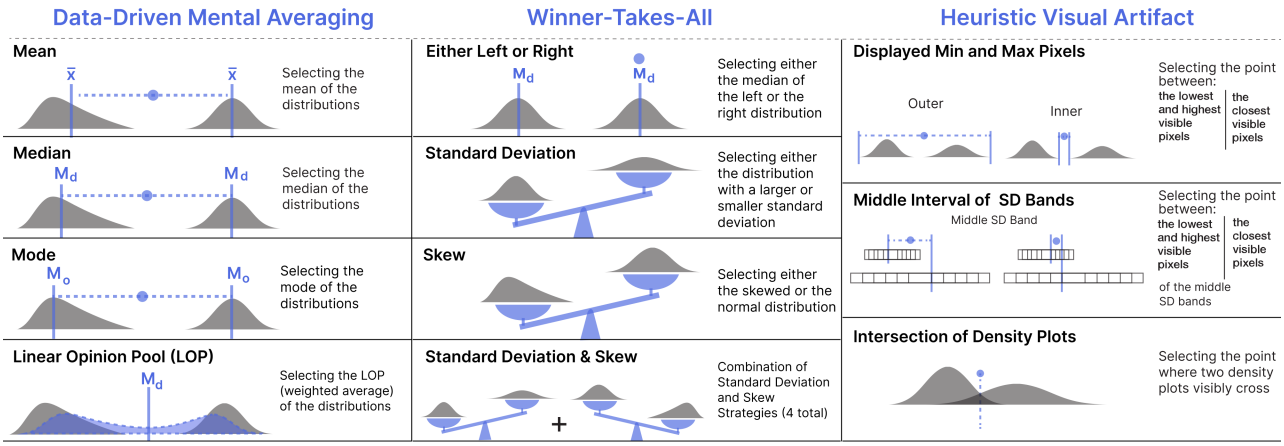


Figure 2: Illustration of the strategy proxies evaluated in our analyses. Strategies are organized into three groups. Data-Driven Mental Averaging (left): participants may mentally combine forecasts by selecting statistical summaries such as the mean, median, or mode of the distributions, or by using a Linear Opinion Pool (LOP). Winner-Takes-All (center): participants may rely on a single distribution, choosing based on standard deviation, skew, or a combination of both. Heuristic Visual Artifacts (right): participants may respond to salient display features, such as selecting the midpoint between the outermost visible pixels, the middle interval of SD bands, or the intersection point of overlapping density plots.

***Winner-Takes-All Strategies**: A strategy in which viewers select one forecast as the only plausible outcome and disregard the others* (Figure 2, center column). Prior work shows that people sometimes discount outlier forecasts in ensembles [48], suggesting that choosing a single forecast can be a natural extension of this behavior. Winner-takes-all decisions also arise in real-world settings. For instance, when two physicians offer conflicting treatment recommendations, averaging their advice is often unreasonable; one must choose which to follow. Similarly, when evaluating competing forecasts, viewers may judge one source as more credible.

Within this class, we identified nine approaches. The most basic selects a forecast without considering its distributional properties (Figure 2, center column, first row). Others rely on simple heuristics, such as choosing the forecast with the smaller or larger standard deviation (Figure 2, second row), reflecting the common misinterpretation that smaller variance implies greater certainty. Additional variants prioritize skew, selecting either the more skewed or more symmetric distribution (Figure 2, third row). The most advanced approaches combine multiple cues, such as selecting the forecast that is both less variable and less skewed (Figure 2, fourth row). To capture this full space, we tested all nine winner-takes-all strategies summarized in the center column of Figure 2.

(5) Randomly selecting the left or right distribution.
(6) Selecting the distribution with a smaller standard deviation.
(7) Selecting the distribution with a larger standard deviation.
(8) Selecting the distribution that is skewed.
(9) Selecting the distribution that is not skewed.
(10) Selecting the distribution with a smaller standard deviation and skew.
(11) Selecting the distribution with a larger standard deviation and skew.
(12) Selecting the distribution with a smaller standard deviation and no skew.
(13) Selecting the distribution with a larger standard deviation and no skew.

***Heuristic Visual Artifact Strategies**: A strategy in which viewers base their judgments on salient visual features of the display rather than on the underlying statistical properties of the forecasts.* Prior work shows that simple visual cues can strongly shape interpretation. For example, McKenzie et al. [35] found that participants were more likely to trust a forecast when it appeared within a sharply bounded 95% CI rather than one with a Gaussian fade. Other studies find that viewers rely on features such as the apparent central tendency, the visual distance between forecasts, or error bar end caps when making judgments [24, 26]. Building on these insights, we reviewed each visualization type to identify heuristic proxies that might arise from its unique visual characteristics, as summarized in the right column of Figure 2.

One strategy that applied across all visualization types involved using the outermost or innermost visible points of the distributions as anchors and mentally averaging those positions (Figure 2, right column, first row). The effectiveness of this heuristic varies by visualization, as the points differ by technique. For instance, a 95% CI is much narrower than SD bands, which span three standard deviations. SD bands also afford a unique variant of this heuristic: because they contain multiple nested intervals, viewers may rely on either the outer edges of the full band or the inner boundaries of the narrowest band (Figure 2, right column, second row). Density plots introduce another distinct heuristic in which viewers focus on the visible intersection of two semitransparent distributions (Figure 2, right column, third row), particularly when the forecasts are close together. The heuristic visual artifact strategies tested are listed below and summarized in the right column of Figure 2:

(14) Averaging the furthest visible points across forecasts.
(15) Averaging the closest visible points across forecasts.

(16) Averaging the closest points of the inner intervals of the SD bands.
(17) Averaging the furthest points of the inner intervals of the SD bands.
(18) Using the intersection point in density plots as the most likely outcome.

In total, we tested 18 distinct strategies across the three categories of Data-Driven Mental Averaging, Winner-Takes-All, and Heuristic Visual Artifact strategies.

3 EXPERIMENTS

To investigate multiple forecast visualization interpretation strategies, we conducted two experiments. The first identified common strategies and examined their variation across visualization types and distribution properties, including normal vs. skewed distributions. The second tested how these findings generalized from two to four forecasts.

3.1 Stimuli Generation

Experiment 1 Datasets: Datasets and stimuli were generated in R [45]. Using the *sn* package v.2.1.1 [1], we sampled 500,000 values to create 40 datasets from normal and skewed-normal distributions with standard deviations of 1 or 2.5. To match the mean and standard deviation of normal counterparts, skewed distributions were generated as follows [37]:

$$\omega_{gen}(\sigma, \alpha) = \sigma \left(1 - \frac{2\alpha^2}{\pi(1+\alpha^2)}\right)^{-\frac{1}{2}}$$

$$\mu_{gen}(\mu, \omega, \alpha) = \mu - \omega \left(\frac{2\alpha^2}{\pi(1+\alpha^2)}\right)^{\frac{1}{2}}$$

Here, μ and σ represent the mean and standard deviation of the normal distribution, α denotes the skewness parameter, and ω represents the scale parameter of the skewed normal distribution. These equations yielded input parameters ω_{gen} and μ_{gen} for the skewed normal function, resulting in skewed distributions with comparable means and standard deviations to the normal ones. To create distribution pairs, we added three units to the mean for forecasts with more agreement and 14 units for those with less agreement (Figure 3a: **agreement low and high**). We generated five combinations of normal and skewed distributions (Figure 3b): both normal (N/N), positive skewed left and normal right (PS/N), positive skewed right and normal left (N/PS), negative skewed left and normal right (NS/N), and negative skewed right and normal left (N/NS), referred to as **shape combinations**. These were crossed with variability levels (Figure 3b), or **SD combinations** (*e.g.,* 1SD/1SD, 2.5SD/2.5SD, 1SD/2.5SD, 2.5SD/1SD), resulting in 40 datasets.

Experiment 1 Stimuli: We selected mark types to test based on (1) the best-performing visualization techniques, (2) appropriate controls, and (3) the most common methods. We plotted two distributions displayed with one of five mark types (median, 95% CI, SD bands, density, and HOPs) in one plot (Figure 1), using the packages *ggplot2* v. 3.4.4 [52] and *ggdist* v. 3.3.1 [29].

We selected density plots and HOPs as two of the most empirically validated techniques for conveying uncertainty [42]. We did not include quantile dot plots, despite their effectiveness, because Experiment 2 required the display of multiple overlapping distributions (*e.g.,* Figure 3a). Although recent methods allow quantile dot plots to visualize two overlapping distributions using multicolored dots [54], this approach does not scale well to more than two distributions. Thus, we chose density plots, better suited for multiple overlapping distributions, and included HOPs, generated by randomly sampling 940 values from our datasets and animating them at 2.5 frames per second. The number of samples and frame rate were configured according to best practices recommended in the seminal work on HOPs [19]. We selected the median mark type as a control without uncertainty, serving as a baseline to assess how people update their decisions when presented with uncertainty.

To evaluate common methods, we included the 95% confidence interval, the most widely used visualization for representing uncertainty (see review on COVID-19 visualizations [57]). We also tested a finer banded-interval version, with each section representing 0.5 standard deviation units, to convey the distribution's shape [54].

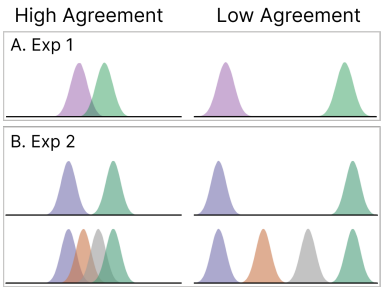
All plot pairs were placed in a single frame to minimize vertical space and avoid scrolling, rather than using small multiples. We colored the distributions purple and green, and created a second set in which we reversed the color encoding to prevent bias. Participants were randomly assigned to one mark-type condition (Median, 95% CI, SD Bands, Density, or HOPs) and one color condition (purple/green or green/purple).

Experiment 2 Datasets: In Experiment 2, we tested only normal distributions, creating stimuli with four and two distributions varying in forecast agreement. Two datasets with four normal distributions were generated, with distributions of greater agreement separated by two units and those with less agreement by six units (see Figure 3a). The procedure of generating normal distributions in Experiment 2 was identical to the procedure in Experiment 1.

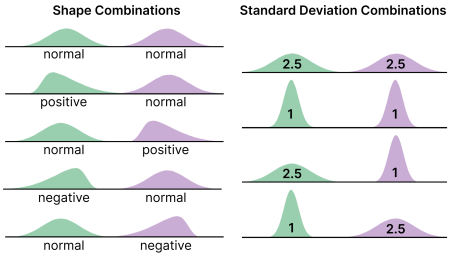
Experiment 2 Stimuli: We tested the same mark types as in Experiment 1. To create two-distribution stimuli, we plotted the outermost distributions from the four-distribution set, maintaining the same overall width (see Figure 3a). The two distributions were colored purple and green, and the four distributions were colored purple, orange, gray, and green, based on ColorBrewer's qualitative color scheme [5], with gray added for color blindness differentiation. Participants did not need to report colors, but they had to distinguish colors in HOPs to correlate the animated samples with their parent distributions.

3.2 Experiment Design

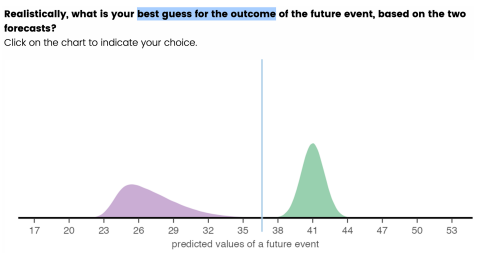
Figure 4 provides an overview of the experimental design and instructions, which were consistent across all experiments. Participants accessed the study online, provided informed consent in accordance with IRB guidelines, and then completed the steps outlined in Figure 4. To enhance the generalizability and minimize context-specific biases (*e.g.,* those that may arise in weather or political forecasting), we employed a neutral scenario. Instructions did not indicate which strategy was correct, nor did participants receive feedback on the accuracy of their responses. This design ensured that participants selected the approach they considered most appropriate, allowing us to examine how forecasts are interpreted with minimal influence from prior knowledge or contextual framing. We revisit this design choice in the limitations section.



(a) High- and low-agreement conditions. Conditions in Experiment 1 (Panel A, two forecasts) and Experiment 2 (Panel B, four forecasts). Density plots show normal distributions with a standard deviation of 1.



(b) Distributional manipulations. The five shape combinations (normal, positive skew, negative skew) are shown on the left. Standard deviation (SD) conditions of 1 and 2.5 are applied across these shapes, illustrated on the right.



(c) Study interface. Screenshot of the task interface used by participants to provide their responses by clicking on the chart to indicate their best guess.

Figure 3: Overview of experimental stimuli and interface. (A) Illustration of conditions showing forecast agreement (high vs. low) and number of forecasts (two vs. four). (B) Distribution shape manipulations and corresponding standard deviation conditions used in the study. (C) Screenshot of the interactive task interface, showing how participants selected their responses.

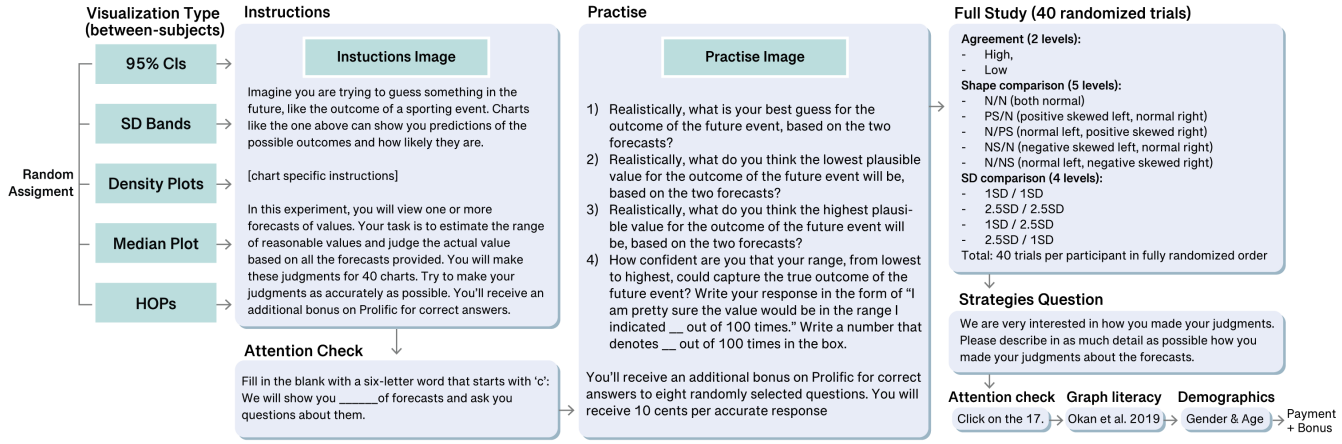


Figure 4: Overview of the experimental design. Participants were randomly assigned to one of five visualization types (95% CIs, SD Bands, Density Plots, Median Plot, or HOPs). They first completed chart-specific instructions, an initial attention check, and a practice trial. The main task consisted of 40 fully randomized trials varying by forecast agreement (high vs. low), shape comparison (five levels), and standard deviation comparison (four levels). After the main task, participants answered an open-ended strategy question, completed a second attention check, graph literacy items, and demographic questions before receiving payment and potential performance bonuses.

Experiment 1: We employed a mixed-subjects design with the following factors: 5 levels of mark type (median, 95% CI, SD bands, density, and HOPs), 2 levels of forecast agreement (low vs. high), 5 levels of shape comparison (N/N, PS/N, N/PS, NS/N, N/NS), and 4 levels of SD comparison (1SD/1SD, 2.5SD/2.5SD, 1SD/2.5SD, and 2.5SD/1SD). This design produced 40 randomized trials per participant, totaling 200 trials across the 5 mark-type groups.

Experiment 2: We examined normal distributions with a standard deviation of 1, using a 5 (mark type: median, 95% CI, SD bands, density, HOPs) x 2 (number of forecasts: 2, 4) x 2 (agreement: low, high) mixed-subjects design. Mark type was the between-subject variable (five groups), and the number of forecasts and agreement were within-subject variables, resulting in four randomized trials per participant, totaling 20 trials across the five mark-type groups.

Instructions: In addition to the instructions shown in Figure 4, participants were provided with chart-specific instructions designed to establish the baseline knowledge needed to interpret each visualization. Because each chart type contained unique visual elements, the instructions varied slightly; however, we edited them to maintain as much consistency as possible across conditions. The full instructions are provided in the supplemental materials and on OSF (osf.io/rwnt8).

Procedure: After reading the instructions, participants completed an attention check that required typing a word beginning

with the letter c to complete the sentence in Figure 4. The intended answer was charts, although any valid c-word (*e.g.,* cats) allowed participants to continue. Those who failed this check were excluded per our preregistered criteria. This approach adheres to Prolific's guidelines, as it assesses instruction following rather than memory. Next, participants completed a practice trial using a forecast not included in the main study. In the main task, participants responded to a validated four-step distribution elicitation measure that we adapted from Speirs-Bridge et al. [49]:

(1) Realistically, what is your best guess for the outcome of the future event, based on the two forecasts?
(2) Realistically, what do you think the lowest plausible value for the outcome of the future event will be, based on the two forecasts?
(3) Realistically, what do you think the highest plausible value for the outcome of the future event will be, based on the two forecasts?
(4) How confident are you that your range, from lowest to highest, could capture the true outcome of the future event? Write your response in the form of "I am pretty sure the value would be in the range I indicated __ out of 100 times." Write a number that denotes __ out of 100 times in the box.

Participants responded by clicking on the visualization to generate a draggable line (Figure 3c). After the initial click, they could adjust their estimate by clicking and dragging the line to update its position or clicking again to redraw the line. The first question appeared alongside the visualization, and Questions 2 and 3 were presented on the next page, where participants marked their lowest and highest plausible values. For Question 4, participants entered a numerical response in a text box.

To encourage engagement, the practice trial informed participants about the possibility of bonus payments (of up to $.80), as detailed in Figure 4. Compensation was determined using a linear opinion pool (LOP; Equation 1) method [2, 50], adapted from [9], with responses within ±0.75 units of the LOP value qualifying. Bonus payments were issued within 24 hours of study completion.

Although recent work on decision-making tasks recommends compensating participants using the same evaluation metrics applied in the study [20], we used the LOP as the mathematically appropriate method for aggregating judgments, since it would have been infeasible to compute bonuses across 18 strategies. Payments were calculated after the study; strategies were not disclosed, no feedback was provided, and each participant could complete the study only once. This ensured the payment structure could not influence responses.

As outlined in Figure 4, after completing the main experiment, participants: (1) responded to an open-ended question about their interpretation strategies, (2) completed a second attention check, (3) completed a graph literacy assessment [38], (4) answered demographic questions, and (5) viewed their bonus.

Promoting Valid Responses: To ensure data quality, we implemented preregistered checks. Eligibility was restricted to participants with a Prolific approval rating of 90% or higher. An attention check was administered at the beginning of the survey, and a final check was required at the end, where participants had to click on a specified number from a chart. These procedures resulted in the exclusion of 21 participants. Although no method guarantees entirely valid responses, combining entry and exit checks with incentive-based participation (*e.g.,* performance bonuses) increases the likelihood of filtering out inattentive respondents while retaining genuinely engaged participants.

3.3 Preregistered Hypotheses

Experiment 1: Our preregistered hypotheses focused on two strategies we expected to be most common. The first was the mathematically ideal method for integrating two distributions: the linear opinion pool (LOP; 1). Greis et al. [17] previously found evidence of participants using this strategy. Likewise, we anticipated that most participants would attempt some form of mental aggregation informed by both the standard deviation and skew of the distributions. For the second, we preregistered hypotheses about the *winner-takes-all* class of strategies, in which participants select one distribution as correct and disregard the other. We predicted that when the distributions were farther apart, participants would be more likely to perceive them as conflicting and adopt a winner-takes-all approach, whereas when they were close together, they would be more likely to average them mentally. This motivated the inclusion of conditions with high agreement (distributions that are close together) and low agreement (distributions that are farther apart). Our preregistered hypotheses focused on these two approaches, but we also explored the broader set of 18 strategies (detailed in subsection 2.1) in an exploratory manner. For Experiment 1, we preregistered the following five hypotheses, ordered to guide the analysis flow:

H1): **When forecasts are close together with uncertainty shown, we predict most participants will mentally average them using a pseudo-variability weighting function. When forecasts differ substantially, we expect more participants to select a value near one forecast rather than between them (winner-takes-all response).** It is well documented that people mentally average multiple expert opinions when expressed numerically (*e.g.,* [6]). However, the tendency to adopt a winner-takes-all approach, where one forecast is deemed reliable and the other rejected, has been less explored, particularly with visualized forecasts. We predicted this behavior based on author disagreements about our personal approaches. Upon reflection, we noted that a high degree of forecast disagreement might suggest one forecast is flawed, whereas general agreement could imply that minor discrepancies explain the differences.

H2: Participants viewing HOPs will be the least likely to show a winner-takes-all response pattern compared to those viewing the other uncertainty visualizations. We predict this because holding multiple distributions in mind and then selecting one is likely to be cognitively demanding when viewing HOPs.

H3: Participants viewing the median mark type will more consistently use a mental-averaging response pattern rather than a winner-takes-all, regardless of the agreement between the forecasts. The median mark type, showing only a single line, completely lacks distributional information, akin to numeric forecasts. Since people typically average numeric forecasts [6], we predicted a similar response for the median mark type.

H4: Participants will update their beliefs toward the skewed distributions. We predicted that skewness would influence judgments, as recent work finds that viewers adjust their decision-making responses based on skewness in ensemble data [48].

H5: When uncertainty is shown with static marks, participants will indicate a range of plausible lowest and highest values by placing lines near the visual ends of each mark type. We predicted that the endpoints of a mark type would bias judgments by acting as visual anchors.

Experiment 2: As a preview, Experiment 1 showed that density plots and SD bands often elicited a winner-takes-all strategy, particularly under low forecast agreement. In Experiment 2, we examined whether this pattern persisted when participants viewed four forecasts. Guided by Gestalt grouping theory, we hypothesized that the presence of four forecasts would strengthen perceptual grouping, even when forecasts were relatively dispersed, thereby increasing the likelihood of mental averaging. Accordingly, we preregistered one hypothesis for Experiment 2 based on this prediction:

H6: Multiple forecast visualizations with four forecasts will elicit a mental-averaging response pattern more frequently compared to those with only two forecasts, which will be more pronounced with high forecast agreement.

3.4 Participants

As specified in our preregistration, we conducted an a priori power analysis using the `pwr` package in R. Following Cohen's guidelines for small-to-medium effects [10], we assumed six degrees of freedom, $\alpha = .05$, power of 0.80, and an effect size of $f^2 = 0.135$, which yielded an estimated requirement of 100.38 participants per group. We rounded this target to 100 and collected data accordingly: 500 participants in Experiment 1 and 500 in Experiment 2. In adherence with our approved IRB recruitment criteria, participants were US-based, over 18 years old, fluent in English, and had an acceptance rate of over 90% on Prolific. Applying attention check and exclusion criteria, we removed 14 participants from Experiment 1, yielding the following condition counts: median (n = 99), 95% CI (n = 98), SD bands (n = 94), density (n = 97), and HOPs (n = 98). In Experiment 2, we excluded seven participants, with condition counts: median (n = 98), 95% CI (n = 95), SD bands (n = 98), density (n = 100), and HOPs (n = 99). Demographics for Experiment 1 were 236 male, 229 female, 16 nonbinary/third gender, 3 prefer not to say, and 3 prefer to self-describe, with a mean age of 39.20 years (SD = 12.54). For Experiment 2, there were 246 males, 238 females, eight nonbinary/third gender, and 1 who preferred not to self-describe, with a mean age of 41.08 years (SD = 14.07).

On average, participants took 34.17 minutes to complete the study. All participants were compensated $8, regardless of completion time, resulting in an average hourly rate of approximately $14. Participants were also eligible for performance-based bonuses, as described in subsection 3.2, up to a maximum of $0.80, and received bonus payments within 24 hours of completing the study.

4 ANALYSIS AND RESULTS

To understand participants' response strategies with multiple forecast visualizations, we conducted a comprehensive exploratory analysis of the strategies influencing participants' judgments. The following section details our analysis approach, followed by results.

4.1 Modeling Approach

For each candidate strategy, we computed the ideal response for each trial and measured the absolute distance (Δ) between participants' responses and these ideals. For example, to evaluate the median strategy, we computed the median of the combined data from both distributions. The median from the combined distributions served as the baseline, and we subtracted it from each participant's response to obtain Δ, representing the difference between the participant's response and the ideal response for that strategy. We calculated these Δ values for all 18 strategies in the units of the displayed distributions.

An artifact of this design means that the potential for error increases when distributions are farther apart. Even if participants attempt to use the median strategy, identifying the midpoint between two widely separated distributions is more difficult than when they are close together. To account for this, we compared the average Δ values for high- and low-agreement conditions under the mean strategy with the median-only visualization. We chose this combination because it produced the lowest raw average Δ overall (Δ = .275), indicating responses in this condition most closely matched the ideal strategy across all visualization types and strategies. The median-only visualization also presents no visual uncertainty, displaying a single line equivalent to two-point forecasts. Prior work shows that people reliably average two numerical forecasts [6], suggesting they would do the same with equivalent visual forecasts. In this case, the average Δ was .205 for high-agreement and .408 for low-agreement conditions. The difference (.203) likely reflects the increased difficulty of selecting the perceptual middle in the low-agreement condition. We therefore subtracted this difference from all low-agreement conditions to adjust for the greater potential error relative to high-agreement conditions.

We then fit a Bayesian multilevel model to predict Δ (absolute deviation from the ideal strategy) with the following specification:

$$\begin{aligned}\Delta_{ij} = \beta_0 &+ \beta_1(\text{Mark Type}_{ij} \times \text{Forecast Agreement}_{ij}) + \\ &\beta_3(\text{Shape Comparison}_{ij}) + \beta_4(\text{SD Comparison}_{ij}) + \\ &\beta_5(\text{Graph Literacy}_{ij}) + u_{\text{ID}[i]} + \epsilon_{ij}\end{aligned} \quad (2)$$

where:

- β_0: Intercept term representing the expected Δ for the reference levels of all predictors.
- β_1: Fixed effect for the interaction between *Mark Type* and *Forecast Agreement*, testing H2 and implicitly accounting for their main effects.
- β_3: Fixed effect of *Shape Comparison*.
- β_4: Fixed effect of *SD Comparison*.
- β_5: Fixed effect of *Graph Literacy*.
- $u_{\text{ID}[i]} \sim \mathcal{N}(0, \sigma_u)$: Random intercept for participant i, capturing within-subject variability.
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$: Residual error term for observation j from participant i.

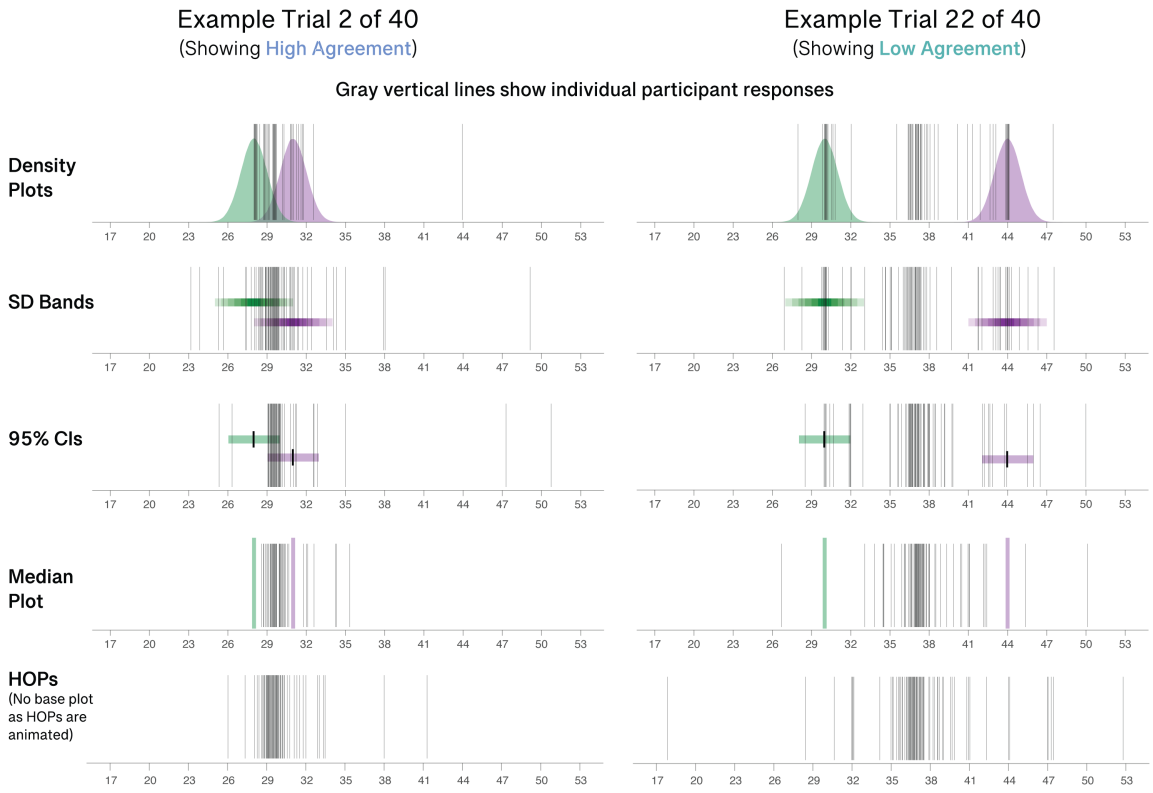


Figure 5: Example participant response patterns for two trials across the five chart types tested. Left column shows a trial with two normal distributions staggered by one standard deviation, exhibiting high forecast agreement. Right column shows examples with low forecast agreement. Solid gray vertical lines represent individual participant responses. These response patterns indicate that Density Plots and SD Bands elicit two primary strategies (mental averaging and winner-takes-all) with different usage across high and low forecast agreement conditions. In contrast, 95% CIs, Median Plots, and HOPs lead to more consistent use of mental averaging. There is also some evidence of a visual heuristic strategy under high forecast agreement conditions for Density Plots (selecting the intersection of density plots) and 95% CIs (averaging the inner points of the CIs).

4.2 Results

All predictors were mean-centered except for Mark Type. Shape Comparison, SD Comparison, and Graph Literacy were included as covariates. Data processing was conducted using the `tidyverse` v. 2.0.0 [53], Bayesian modeling was performed using `brms` v. 2.20.4 [7], posterior summaries were obtained with `tidybayes` v. 3.0.6 [28], and visualizations were created with `ggdist` v. 3.3.1 [29].

Priors. We used weakly informative priors:

$$\beta_k \sim \mathcal{N}(0.5, 2.5) \quad \text{for all fixed effects}$$
$$\sigma_u \sim \mathcal{N}(0, 2.5) \quad \text{for the random intercept SD}$$

Analysis Scope: We fit 18 models, 15 using the full dataset and the remaining models on visualization-specific subsets. This approach enabled us to rank strategies by their Δ values, identifying those that were most closely aligned with participant responses. We did not directly compare the models using fit indices; instead, we describe the relationships between the models qualitatively. Within each model, direct comparisons between predictors are appropriate. All "CIs" reported refer to 95% credible intervals.

To illustrate common response patterns, we plotted participants' actual responses for two representative stimuli in Figure 5. The vertical lines show participant responses overlaid on the stimuli. Visual inspection reveals two predominant patterns. In the first, participants placed lines near the modes of the individual distributions, effectively selecting one distribution over the other, a winner-takes-all strategy. This pattern is most apparent for density plots and SD bands under low forecast agreement (right column of Figure 5). The second, and more common, pattern reflects a mental averaging strategy, with participants placing lines near the central tendency of the two distributions. This strategy is especially prominent for 95% CIs, mean plots, and HOPs in both the high- and low-agreement examples. In the subsequent analysis, we evaluate the statistical robustness of these observed patterns using the Bayesian models described in subsection 4.1.

Following recommendations for transparent reporting of Bayesian analyses [31], we report posterior distributions for model parameters, including posterior means, 95% credible intervals, and the probability of direction (the proportion of the posterior strictly above or below zero). We present the posteriors in visualizations

with a consistent y-axis, where zero indicates no meaningful effect, and the degree of displacement from zero reflects the effect strength, allowing results to be interpreted in terms of effect magnitude and associated uncertainty rather than threshold-based significance. For reference, the x-axis scale of all stimuli spanned 36 units; if a participant ignored the visualization and clicked at the furthest extent of the scale, the resulting difference would be 36 units subtracted from the reference strategy's measure. We report the Δ values on this same scale for ease of interpretation. In the following section, we first evaluate each hypothesis and then present additional insights from analyzing all strategies, followed by descriptive results on participants' self-reported strategies.

4.2.1 Findings H1. ***When forecasts are close together and uncertainty is shown, we predict that most participants will mentally average them using a pseudo-variability weighting function. When forecasts differ substantially, we expect more participants to adopt a winner-takes-all response.*** To test H1, we compared the LOP (variability weighting function) strategy with a winner-takes-all strategy (see panels A and B of Figure 6). Among the winner-takes-all strategies, comparing participants' responses to *selecting the median of forecast A or B* (Δ = 2.76) yielded the lowest Δ values across all conditions. Therefore, we used *selecting forecast A or B* as a representative winner-takes-all approach to compare to the LOP.

As shown in Panel A of Figure 6, for the low forecast agreement conditions, participants' responses when viewing Density Plots (Δ = 2.22) and SD Bands (Δ = 3.32) were more in line with the winner-takes-all strategy (selecting forecast A or B) compared to the response patterns of those who viewed HOPs (Δ = 5.33), 95% CIs (Δ = 4.67), and Median Plots (Δ = 5.58). Furthermore, under low forecast agreement conditions, participants who viewed Density Plots and SD Bands gave responses that were more indicative of the winner-takes-all strategy (selecting forecast A or B) than the LOP strategy, an effect evident in the higher Δs for Density Plots and SD Bands when comparing Panel A to Panel B in Figure 6. For Density Plots, participants' responses were more indicative of the winner-takes-all strategy (selecting forecast A or B; median posterior Δ = 2.22, 95% CI = [1.99, 2.44]) than the LOP strategy (median posterior Δ = 3.77, 95% CI = [3.50, 4.04]). For SD Bands, participants' responses showed a similar pattern, being more consistent with the winner-takes-all strategy (selecting forecast A or B; Δ = 3.32, 95% CI = [3.09, 3.55]) than with the LOP strategy (Δ = 4.73, 95% CI = [4.46, 5.01]).

These results indicate that, under low forecast agreement conditions, participant responses to Density Plots and SD Bands were more consistent with the winner-takes-all strategy (selecting either forecast A or B) than a weighted mental-averaging strategy (LOP). In contrast, under low forecast agreement conditions, participants who viewed HOPs and 95% CIs gave responses that were more indicative of the LOP strategy. Therefore, **H1 is confirmed only for Density Plots and SD Bands**: Although this finding does not align with our hypothesis, it is noteworthy that HOPs and 95% CIs promoted a strategy that theorists may consider mathematically ideal, even when forecasts were far apart. Designers seeking to encourage a mental averaging strategy may therefore prefer using HOPs or 95% CIs over Density Plots or SD Bands.

Another critical, unanticipated finding emerged when examining mental averaging strategies beyond the LOP. Across all 18 strategies tested, the most elementary approach—mentally computing the mean of the two distributions—yielded the lowest Δ value (average Δ across all conditions = 1.68) consistently. For comparison, the average Δ across all conditions for the LOP strategy was 2.76. As shown in panel C of Figure 6, participants' responses were nearly always more consistent with the mean strategy, yielding equal or lower Δ values compared to the LOP (panel B). The one exception occurred for Density Plots under low forecast agreement conditions, where participants' responses were more indicative of the LOP strategy (panel B, Δ = 3.77) than the mean strategy (panel C, Δ = 4.35). Overall, these results provide evidence that participants' responses were sometimes more indicative of a simple mental averaging strategy (*e.g.,* averaging the means) rather than a weighting function that incorporated standard deviation or skew. This finding contrasts with prior work suggesting that people tend to use a weighted average strategy [17]. Our results may diverge from past work because we tested a substantially larger set of strategies, allowing for consideration of a broader range of alternatives.

4.2.2 Findings H2: ***Participants viewing HOPs will be the least likely to show a winner-takes-all response pattern compared to those viewing the other uncertainty visualizations.*** To test H2, we examined the posteriors shown in Panel A of Figure 6. Under low forecast agreement conditions, participants' responses when viewing HOPs were less indicative of a winner-takes-all strategy (selecting forecast A or B), yielding substantially larger posterior Δ values (Δ = 5.33, 95% CI = [5.11, 5.56]) compared to responses from participants viewing Density Plots (Δ = 2.22, CI = [1.99, 2.44]), SD Bands (Δ = 3.32, CI = [3.09, 3.55]), and 95% CIs (Δ = 4.67, CI = [4.45, 4.91]). However, this pattern did not persist under high forecast agreement conditions, where participants' responses showed no meaningful differences across visualization types. This analysis **partially supports H2**, indicating that under low forecast agreement conditions, participants' responses to HOPs were less indicative of a winner-takes-all strategy than responses to other uncertainty visualizations.

4.2.3 Findings H3: ***Participants viewing the median mark type will more consistently produce a mental-averaging response pattern rather than a winner-takes-all response pattern, regardless of the agreement between the forecasts.*** To test H3, we plotted the posterior results of all of the models, grouped by visualization type, in Figure 7. Panel E in Figure 7 presents the results for the Median Plot, comparing the posterior of the mental-averaging strategy (selecting the mean of the two forecasts) with the winner-takes-all strategy (choosing the median of forecast A or B). These were the only two strategies applicable to the Median Plot, since it did not convey distributional information and therefore could not be fairly evaluated using strategies that required it. The analysis shows that participants who viewed the Median Plot gave responses more indicative of the mental averaging strategy (average Δ = 0.67) than the winner-takes-all strategy (selecting the median of forecast A or B; average Δ = 3.51), **confirming H3**. This finding aligns with prior work showing that people mentally average point-based numerical forecasts [6], and it provides evidence that these results generalize to the visualization formats and context tested here.

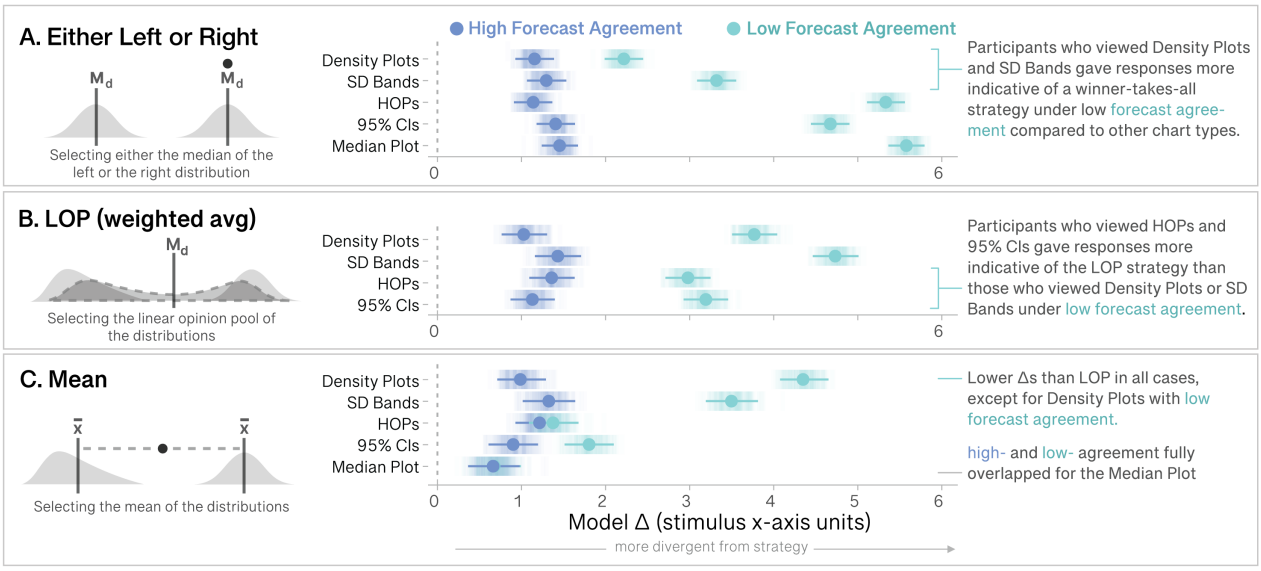


Figure 6: Posterior model results (Δ) showing three strategies of the 18 analyzed. (A) Winner-takes-all: participants select one of the forecasts (A or B). (B) Linear Opinion Pool (LOP; weighted average). (C) Mean averaging of forecasts. Across visualization types (Density Plots, SD Bands, HOPs, 95% CIs, Median Plots), points indicate posterior means, horizontal intervals denote 95% credible intervals, and vertical lines show posterior samples. Blue = high forecast agreement, teal = low forecast agreement.

4.2.4 Findings H4: ***Participants will update their beliefs toward the skewed distributions.*** We examined the impact of skew under both the mental averaging and heuristic visual artifact models, as well as the winner-takes-all models. For mental averaging and heuristic visual artifacts, skew was assessed through the *shape comparison* term in our hierarchical model (see Equation 2). We included this term to test whether differences between skewed and normal distributions accounted for meaningful variation in participants' responses. Across all models, we found no meaningful evidence that skew accounted for variation in participants' responses, as every 95% credible interval included zero. This null result held consistently for both the mental averaging and heuristic visual artifact models.

To evaluate skew in the winner-takes-all models, we focused on strategies that explicitly incorporated skew. Using Figure 7, we compared participants' responses to idealized strategies in which they would have consistently selected the skewed distribution when paired with a normal distribution, as well as strategies in which they would have consistently selected the skewed distribution combined with either a smaller or larger standard deviation. For HOPs, SD Bands, and 95% CIs, participants' responses compared to the idealized strategy of selecting the skewed distribution never yielded lower Δ values than the corresponding strategy of choosing the normal distribution. The sole exception occurred with Density Plots, in which participants' responses were more consistent with the idealized strategy of selecting the skewed distribution (average high and low forecast agreement $\Delta = 3.77$, CI = [1.56, 5.95]; Figure 7 Panel A, row **Select Skewed**) than with selecting the normal distribution (average high and low forecast agreement $\Delta = 4.61$, CI = [1.64, 7.60]; Figure 7 Panel A, row **Select Normal**). However, when comparing strategies such as choosing the Normal + Smaller SD (average high and low forecast agreement $\Delta = 2.43$, CI = [0.95, 3.97]) versus the Skewed + Smaller SD (average high and low forecast agreement $\Delta = 2.58$, CI = [1.03, 4.12]), participants' responses showed no meaningful differences.

Overall, these results provide no compelling evidence that participants systematically incorporated skew into their decision-making. We therefore **reject H4**.

4.2.5 Exploratory follow-up: Before turning to H5, which concerns the second part of the task in which participants placed lines to indicate the highest and lowest plausible values, we will discuss the complete set of strategies, including those not directly tied to our hypotheses.

Two strategies were relatively consistent with participants' responses in some cases: one associated with Density Plots and the other with SD Bands. For Density Plots (Figure 7, Panel A first row, **Intersection**), a strategy linked to relatively small Δ values involved using the point of intersection between the two distributions as an indicator of the most likely value. Under high forecast agreement, this strategy was among the most consistent with participants' responses ($\Delta = 0.89$, CI = [0.67, 1.10]). Under low agreement, however, Δ values increased substantially, driven primarily by cases in which two broad distributions (each spanning three standard deviations) produced almost imperceptible visual overlap. Furthermore, we observed a second visualization-specific strategy with SD Bands (Figure 7 Panel B, fourth row **Avg. Center Interval Inner**). If followed, this strategy would involve using the center interval and then averaging the inner points of that interval (illustrated in Figure 2, middle of the Heuristic Visual Artifact column). The Average Center Interval Inner strategy was the fourth most aligned with

Figure 7: Posterior model results (Δ) for 18 candidate strategies across five visualization types: (A) Density Plots, (B) SD Bands, (C) HOPs, (D) 95% CIs, and (E) Median Plots. Blue = high forecast agreement; teal = low forecast agreement. Points indicate posterior means, horizontal intervals denote 95% credible intervals, and faint vertical lines show posterior samples.

participants' responses for SD Bands, yielding relatively small Δ values under both high ($\Delta = 1.34$, CI = [1.01, 1.67]) and low forecast agreement ($\Delta = 3.55$, CI = [3.23, 3.88]). Both of these strategies were relatively highly aligned with participants' responses and relied on salient visual artifacts that varied across visualizations, highlighting how heuristic features can guide judgments even without a strong normative basis.

Across visualization types, participants' responses were also aligned with strategies that focused on the outer extent of intervals, particularly averaging the closest pair of the furthest visible points (labeled **Avg. Inner Points** in Figure 7). This strategy was consistently aligned with participants' responses across conditions and was especially prominent for HOPs (averaged across high and low agreement $\Delta = 1.54$, CI = [1.19, 1.86]) and 95% CIs (averaged across high and low agreement $\Delta = 1.69$, CI = [1.10, 2.28]), ranking fourth overall. For HOPs, participants' responses were particularly consistent with a process of mentally retaining the closest samples from each distribution and comparing those values. Such a simple heuristic may reduce the cognitive effort required to represent the full distributions, which may explain why it is highly aligned with responses for HOPs in both high- and low-agreement conditions.

Taken together, these exploratory analyses indicate that participants' responses were aligned with a wide range of strategies, many of which diverged from normative approaches. Importantly, this variation highlights that focusing exclusively on idealized strategies risks overlooking the heuristics and alternative approaches that people may use when interpreting forecasts.

4.2.6 Findings H5: ***When uncertainty is shown with static marks, participants will indicate a range of plausible lowest and highest values by placing lines near the visual ends of each mark type.*** In the second part of the task, participants placed lines to indicate where they believed the outer extent of the forecast would be. To illustrate this behavior, we plotted the raw responses directly on the stimuli for a representative trial, showing where participants positioned the upper and lower bounds (see Panel A of Figure 8). Visual inspection suggests that participants frequently placed these endpoints at or near the visible extremes of the distributions for 95% CIs, SD Bands, and Density plots.

Interestingly, participants' responses to Median Plots, which lack explicit uncertainty, indicated that they did not consistently place the upper and lower bounds at the edges of the median mark (see Panel A of Figure 8). Instead, participants appeared to extend the bounds beyond the mark itself. This pattern reinforces prior findings [24] that people intuitively infer uncertainty in forecasts

even when it is not explicitly depicted. However, the way participants mentally added this uncertainty was highly variable. Another notable result was that participants' responses to HOPs aligned with those of Density Plots (Figure 8 Panel A), demonstrating their ability to extract complex statistical information from animated uncertainty representations. Importantly, we did not control viewing duration, meaning that even with an uncontrolled number of samples, participants tended to construct interpretations of the distributions that closely resembled Density Plots.

To formally test H5, we coded responses within 1.5 units (45 pixels) of the maximum visual boundary as 1 (*i.e.,* selecting the visual boundary). Because HOPs were animated and their perceived range varied with viewing time, they were excluded from this analysis. We then conducted a preregistered binomial logistic multilevel Bayesian regression. For parsimony, we fit a single model rather than separate models for upper and lower bounds, deviating slightly from our preregistered plan by including a predictor for boundary type (Upper/Lower Boundary). As with the other models, we specified uninformative priors. The model was defined as follows, using the same variable definitions as in Equation 2:

$$\begin{aligned}\text{Vis Ext}_{ij} = \beta_0 &+ \beta_1(\text{Mark Type}_{ij}) + \beta_2(\text{Forecast Agreement}_{ij}) + \\ &\beta_3(\text{Shape Comparison}_{ij}) + \beta_4(\text{SD Comparison}_{ij}) + \\ &\beta_5(\text{Upper/Lower Boundary}_{ij}) + \beta_6(\text{Graph Literacy}_{ij}) \\ &+ u_{\text{ID}[i]} + \epsilon_{ij}\end{aligned} \tag{3}$$

Panel B of Figure 8 shows the posterior distributions from our analysis, where the x-axis represents the probability of placing lines at the visual extent of each mark type. This analysis reveals that participants' responses to 95% CIs and SD Bands were most often aligned with strategies that placed boundaries at the visual extents, likely because these mark types provide the most distinct visual boundaries. In contrast, only 28–38% of participants' responses to Density Plots placed boundaries at the extremes, likely reflecting the difficulty of discerning their full extent. This pattern persisted regardless of forecast agreement. We therefore find evidence to **accept H5**.

4.2.7 Self-Reported Strategies. Two trained coders independently reviewed the 500 free-response explanations from Experiment 1, applying the three a priori strategy classes (*data-driven mental averaging*, *winner-takes-all*, *heuristic visual artifact*) and using open coding to capture any additional unanticipated strategies. Because strategies were not mutually exclusive, responses could receive multiple codes. After the initial pass, coders cross-checked one another's work and corrected occasional errors. Figure 9 shows the proportion of responses assigned to each strategy class (averaged across coders). Inter-rater reliability was high for all categories: *mental averaging* ($\kappa = .94$, $z = 20.70$, $p < .001$), *winner-takes-all* ($\kappa = .91$, $z = 20.10$, $p < .001$), and *heuristic visual artifact* ($\kappa = .90$, $z = 19.80$, $p < .001$). One coder then conducted a second pass to examine whether finer-grained subcategories could be reliably identified. Most responses lacked sufficient detail for subcoding, although we report proportions where subcategories could be inferred.

Because many participants described only one aspect of the task, our primary analyses focused on coding the main task. However,

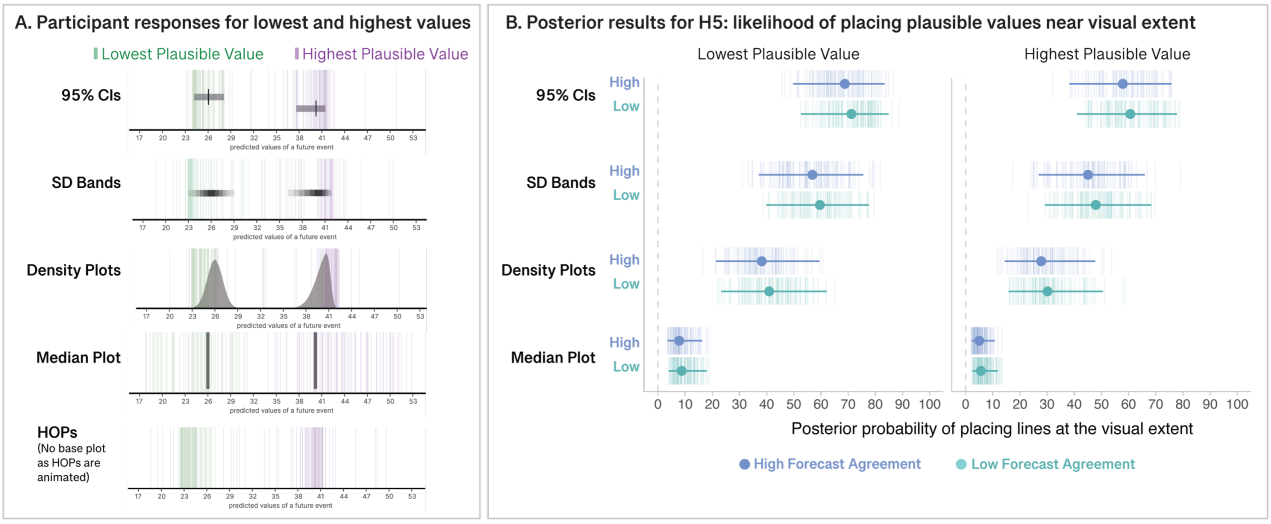


Figure 8: (A) Participant placements of lowest (green) and highest (purple) plausible values across mark types for a representative trial. HOPs have no static base plot, but participant responses are included. (B) Posterior estimates of the probability that these values were placed within 1.5 units of the visual extent, with blue = high forecast agreement stimuli and teal = low forecast agreement stimuli. Points show posterior means, horizontal intervals denote 95% credible intervals, and faint vertical lines indicate 200 posterior samples per model. Together, the two panels show that participants commonly placed endpoints at or near the visible extremes of the distributions for 95% CIs, SD Bands, and Density Plots, but not for Median Plots.

an unanticipated strategy also emerged from the task where participants reported the highest and lowest plausible values, which we describe in the final section.

Data-Driven Mental Averaging Strategies: The qualitative coding revealed patterns that closely aligned with the quantitative results. As shown in Figure 9, participants who viewed the Median, 95% CI, and HOPs displays were more likely to report strategies consistent with mental averaging than those who viewed Density Plots or SD Bands. For example, one HOPs viewer explained, "When making my best estimate on what the forecast would be, I opted to pick a value in the middle of the two medians because I thought that would be the most probable range."

Within this strategy class, participants described two approaches. Some attempted a mathematical operation, such as estimating a statistical moment (*e.g.,* "I roughly averaged the means," "I picked as close to the median of the two bars as possible"; 14.2%). Others relied on a broader visual judgment (*e.g.,* "I would usually look for the middle and use that to gauge what I was going to pick"; 29.2%), reflecting a graph-reading process rather than a formal calculation. A smaller subset (7.8%) reported more sophisticated methods that incorporated distributional features such as skew or variability. For example, one participant explained, "I made my average prediction based on what I considered to be the 'average' of the two forecasts. If the forecasts were overlapping, I used how they were weighted (perfect bell or biased left/right) and which had the higher vertical value, and selected a value that would be the average of both bells combined." Weighted mental averaging (LOP) was most common for Density Plots (23.68%), followed by the 95% CI (21.05%), HOPs (18.42%), Median Plots (18.42%), and SD Bands (18.42%).

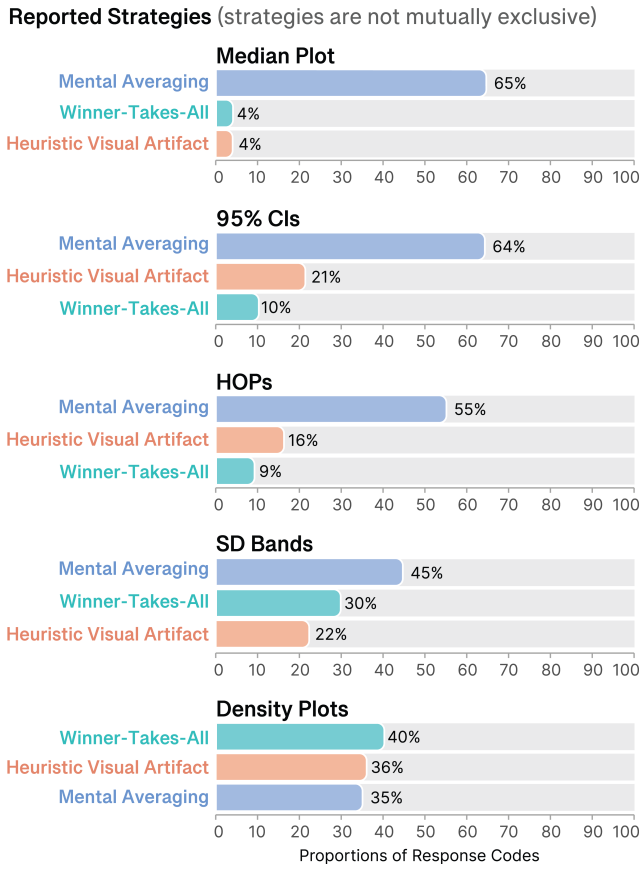


Figure 9: Proportion of responses coded into each strategy class (Mental Averaging, Winner-Takes-All, and Heuristic Visual Artifact) by visualization type. Responses were coded by independent raters, and categories were not mutually exclusive, allowing participants to report multiple strategies. Participants viewing Median Plots, 95% CIs, and HOPs most frequently reported mental averaging, whereas responses for SD Bands and Density Plots showed a more mixed pattern. Density Plots were most strongly associated with winner-takes-all responses.

Winner-Takes-All Strategies: The coded winner-takes-all strategies closely aligned with the quantitative findings as well. As shown in Figure 9, participants who viewed Density Plots (40%) or SD Bands (30%) were more likely to report winner-takes-all approaches than those viewing 95% CIs (10%), HOPs (9%), or Median Plots (4%). Fifty-two percent of participants who used this strategy reported selecting the forecast with the smaller standard deviation. For example: "*...For charts where the two lines didn't intersect, I just chose wherever the peak was highest, and if the two peaks were equally high, I randomly chose one.*" Another participant noted, "*...I was choosing which line appeared to have better data (less variance) and choosing that as the most likely...*" Although participants often selected the forecast with the smaller SD, their explanations typically referenced the height of the mode (*i.e.,* choosing the "tallest" or "highest" peak), rather than explicitly focusing on spread.

Heuristic Visual Artifact Strategies: For the heuristic visual artifact class, participants' written explanations generally lacked the specificity needed to distinguish substrategies at the same level of granularity as the quantitative analysis. Even so, several recurring patterns aligned closely with the quantitative findings. A common strategy across all visualization types except the Median Plot (as this mark was only point-based) involved using the visible extent of the forecasts to form an approximate average. This approach was most frequently reported for HOPs (41.18%), followed by Density Plots (23.53%), 95% CIs (17.65%), and SD Bands (5.88%). These results mirror the quantitative patterns, in which averaging the inner and outer bounds showed some of the lowest deltas for HOPs in both high and low disagreement conditions. As one participant described, "*To make my judgments, I carefully watched the moving lines and tried my best to memorize the highs and lows...For my estimated guess of what I thought the forecast would be, I tried to use an approximate mental average between the highs and lows.*"

For Density Plots, the most common heuristic visual artifact strategy involved using the point where the two distributions intersected as the most plausible outcome. As one participant explained, "*...When the two forecasts overlapped, I assumed that their point of intersection was the best guess for the outcome of the event....*" Among Density Plot viewers who reported a heuristic visual artifact strategy, 54.29% used this intersection heuristic, followed by 11.42% who relied on the visible extent of the distribution. This pattern closely matches the quantitative results, which also identified the intersection heuristic as the dominant strategy for Density Plots.

For SD Bands, participants most often anchored their judgments on the inner interval of the bands. This included averaging the inner or outer intervals of both bands or selecting their midpoints. One participant noted, "*...if both of the forecasts were within each other's range, the plausible values I would put at the beginning and ending parts of the smallest range.*" Among participants whose heuristic sub-strategy could be inferred, 47.62% reported using this inner-band heuristic, with the next most common approach being reliance on the visible extent of the mark (4.76%).

Strategy Switching: Participants frequently reported switching strategies depending on whether forecasts showed high or low agreement. The likelihood of using multiple strategies varied by visualization type, with SD Bands (40%) and Density Plots (39%) showing the highest rates, followed by 95% CIs (26%), HOPs (21%), and Median Plots (7%). Participants described this switching explicitly. One participant noted, "*When the 2 forecasts overlapped, I assumed that their point of intersection was the best guess for the outcome of the event. When the 2 forecasts did not overlap, I chose the higher peak of the 2 forecasts.*" Another explained, "*If the two predictions were very different I was thinking that one of the predictions is completely wrong and should be discounted then it was a matter of just picking one of the very different predictions, I tended to prefer a more narrow guess of ranges... If the two choices were very close then I took the average of the two spectrums since they both kind of reinforced the legitimacy of the other prediction.*" This pattern corroborates our quantitative findings: Density Plots and SD Bands were most likely to elicit a winner-takes-all approach when agreement was low, but participants often switched to mental averaging when agreement was high.

Strategy for High/Low Plausible Values: Open coding revealed two main approaches participants used when selecting the highest and lowest plausible values. Because this task was not a core focus of the study, no *a priori* codes were specified; however, two strategies emerged consistently. The most common approach aligned with our prediction that participants would rely on the visible extent of the mark, placing their upper and lower bounds at the furthest visible points in the display. A substantial majority of relevant responses (81.13%) used this straightforward method, while others applied a modified version in which they located the visual endpoint and then added a small buffer beyond it or slightly inside it. For example, one participant wrote, "*For the highest plausible value, I picked the highest number I saw and maybe a tad higher.*"

A second, unanticipated strategy reflected a more statistical intuition (14.15% of relevant responses). Here, participants began with their predicted value and constructed a range around it that resembled a confidence interval. They implemented this in varied ways, such as adding fixed amounts above and below their estimate or explicitly attempting to approximate a 95% interval. Illustrative examples include, "*I went about in the middle for the initial guess. Then I went about 10% or so above and below for the higher and lower guesses.*" and "*I tried to choose 95% confidence intervals...I tried to estimate about 2.5% outside of the confidence interval (tails outside the lowest and highest values).*" These descriptions suggest two overarching perspectives: most participants focused on visually capturing the full extent of the forecast distribution, while a smaller subset constructed an interval around their own predicted value.

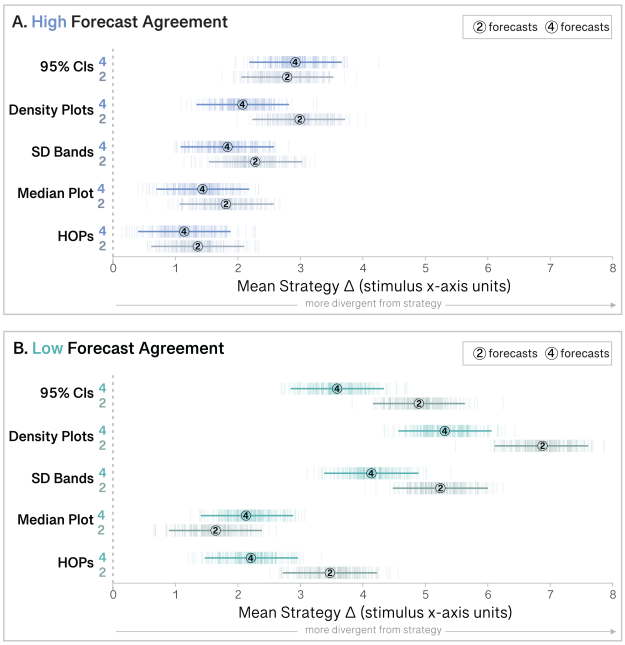


Figure 10: Posterior distributions for H6, the difference between responses and the mean mental averaging strategy. Panel A shows results under high forecast agreement and Panel B under low forecast agreement, across two and four forecasts. Circles mark posterior means, horizontal lines show 95% credible intervals, and vertical lines represent 200 posterior samples. Under low forecast agreement (Panel B), increasing the number of forecasts from two to four meaningfully shifted responses toward the mean strategy (reduced the Δ) for visualization types that explicitly encode uncertainty (95% CIs, Density Plots, SD Bands, and HOPs). Increasing the number of forecasts had no meaningful effects in other conditions.

4.2.8 *Findings H6:* ***Multiple forecast visualizations with four forecasts will exhibit a greater occurrence of the mental- averaging response pattern compared to those with only two forecasts, which will be more pronounced with high forecast agreement.***

To test this hypothesis, we applied the same Bayesian model and data processing procedures described in Equation 2, but using data from Experiment 2, which included four forecasts. We focused specifically on the mean mental averaging strategy, as it was among the most common in Experiment 1 and aligned with our preregistered hypothesis. Figure 10 presents samples from the Bayesian posterior distributions, organized by forecast agreement (left panel high agreement and right panel low agreement). For each mark type, results from conditions with four forecasts and two forecasts are plotted on the same line to facilitate comparison.

Our analysis provides partial support for the hypothesis that forecasts with four distributions would make participants' responses more indicative of a mental averaging approach, but only under low forecast agreement and for mark types that incorporated uncertainty. As shown in Figure 10, the number of forecasts did not

influence participants' responses to the Median Plot under either high or low forecast agreement. However, under low forecast agreement, participants' responses to 95% CIs, HOPs, SD Bands, and Density Plots showed meaningful separation in the posterior distributions between two- and four-forecast conditions. Thus, **H6 is only partially supported**.

5 DISCUSSION AND FUTURE WORK

In this work, we conducted two carefully controlled experiments using empirically validated elicitation measures to examine how people interpret multiple forecast visualizations and report their best guesses of future outcomes. Our aim was to isolate the perceptual and cognitive effects of the visualization techniques themselves. To do so, we adopted a context-free design that establishes a baseline understanding of how people interpret multiple forecasts in the absence of domain cues. Domain-specific scenarios (*e.g.,* weather, finance, health) can introduce strong biases and idiosyncratic reasoning that obscure these underlying perceptual effects. This baseline is therefore essential: without it, disentangling whether observed behaviors arise from visualization designs or from domain-specific beliefs is difficult. The foundational patterns we uncover in this paper will enable forecasters and designers to anticipate a baseline with which their forecasting contexts might interact, amplify, or attenuate. Throughout both studies, we identified clear quantitative patterns and explicit participant descriptions indicating several distinct strategies, which shifted systematically depending on the visualization type and the properties of the forecasts.

Consistent with prior research on text-based forecasts [6], participants' responses were often indicative of a mental integration strategy, particularly when interpreting 95% confidence intervals, median forecasts, and HOPs. A substantial proportion of participants also directly reported attempting to mentally average the forecasts. At the same time, a sizable proportion of participants' responses were aligned with a winner-takes-all approach, in which participants judged one forecast as more likely rather than integrating across forecasts. This tendency was especially pronounced for density plots and SD bands under low forecast agreement. In addition, participants' responses were also aligned with heuristic strategies driven by salient visual artifacts, such as focusing on crossover points or end caps, consistent with prior work on visual proxy strategies [23, 26, 39, 55]. These strategies were more common when such artifacts were visually prominent.

A key contribution of this work is the examination of the lesser studied winner-takes-all strategy. One participant explained this reasoning clearly: "*If the two predictions were very different I was thinking that one of the predictions is completely wrong and should be discounted then it was a matter of just picking one of the very different predictions...*" A recent 50-year review details several approaches for down-weighting or excluding poorly performing forecasts [51]. It is plausible that some members of the general public intuitively adopt a similar logic. At the same time, we remain neutral regarding the "correctness" of any particular response strategy. Forecasting scholars consistently emphasize that no single combination method is universally optimal (*e.g.,* [9, 51]), and that the appropriate approach depends on factors such as data structure, model quality, diversity among model assumptions, and the specific forecasting problem at hand [51]. Given this lack of consensus, it is essential to document how visualization techniques afford some strategies more readily than others [15], thereby providing forecasters and designers with evidence to guide their selection of visualization techniques that best support their intended goals.

In analyzing 18 possible strategies, we identified a subset of conflicting approaches in which many participants relied on perceptual artifacts of the visualizations rather than underlying distributional information. These strategies include identifying the intersection points of density plots, focusing on the endpoints of 95% CIs, and using the center bands of standard deviation intervals to approximate central tendencies (see Figure 2, right for example illustrations). These visual heuristic strategies align with prior findings that viewers often depend on visualization proxies to infer values [23, 39, 55]. These heuristics pose issues because the visual features on which the participants focus vary across different visualizations, introducing significant unintended variability in the way people reason about multiple forecasts.

We also observed evidence of a visual-proxy–based deterministic construal error in participants' responses when interpreting forecasts' range. Responses often aligned the upper and lower boundaries with the visual extremities, particularly for mark types with clearly defined edges. This pattern suggests a bias toward boundary anchoring or the use of endcaps as proxies for data limits. Notably, participants rarely placed upper and lower bounds near the visualized line of the median plot, consistent with prior research showing that people infer uncertainty even when it is not explicitly conveyed [24]. These findings provide concrete evidence to illustrate the broad range of additional uncertainty participants applied to forecasts.

Based on the results of this work, we provide the following design recommendations for designers who wish to use multiple forecast visualizations to communicate competing forecasts. These suggestions highlight how visualization choices can shape users' interpretation and integration of forecast information:

- **Align visualization types with desired strategies.** Our results showed that HOPs, 95% CIs, and median plots consistently encouraged mental averaging strategies in both high- and low-agreement conditions. In contrast, density plots and SD bands led participants to alternate between mental averaging and a winner-takes-all strategy depending on the level of agreement between forecasts (Sections 4.2.1, 4.2.2, and 4.2.3). Designers seeking to consistently promote mental averaging of forecasts should use HOPs, 95% CIs, or mean plots. Likewise, the LOP strategy (*i.e.,* performing a mental weighting function that incorporates differences in standard deviation and skew) is more likely to be evoked when using HOPs and 95% CIs under high forecast agreement conditions than density plots and SD bands. This strategy represents the most statistically advanced method we considered for mentally averaging forecasts.
- **Use more forecasts to encourage mental averaging.** When participants viewed four forecasts, they were more likely to adopt mental averaging strategies compared to two-forecast conditions (Section 4.2.8). This finding suggests that increasing the number of forecasts displayed can nudge users

toward a mental averaging technique rather than a winner-takes-all approach.

- **Anticipate winner-takes-all strategies.** Density plots and SD bands led participants to adopt a winner-takes-all strategy more often under low agreement, selecting a single distribution rather than integrating across forecasts (Section 4.2.1). Designers should be aware of this tendency and consider providing additional information to support users who adopt a winner-takes-all approach. For example, presenting metadata about forecasters or model properties (*e.g.,* historical accuracy or reliability) could help users more effectively calibrate which forecast to prioritize. We found that participants often preferred the forecast with the smallest standard deviation, suggesting that they may be inclined to use contextualizing information about forecast properties when applying this strategy.
- **Account for perceptual heuristics.** Participants frequently relied on visual artifacts such as intersections in density plots or inner bands in SD bands when making judgments (Section 4.2.5). These perceptual heuristics strongly influenced their interpretations.
- **Be aware that mark span shapes perceived uncertainty.** In the range-elicitation task, participants often assumed that the extent of possible outcomes ended near the farthest visible extent of the mark (Section 4.2.6). For example, when viewing SD bands, participants inferred a broader range of outcomes than when viewing 95% CIs, indicating that the span of the mark strongly shaped their assumptions about uncertainty. Designers should be aware of this tendency and the inconsistency it introduces across visualizations. Interestingly, a smaller subset of participants reported attempting to mentally construct their own confidence intervals. If applied consistently, this strategy would be less dependent on the visual extent of the mark. Future work could explore whether prompting or supporting this kind of interval-based reasoning improves the reliability of range judgments.
- **Make uncertainty explicit.** We also found that when no uncertainty information was shown (*e.g.,* median plots), participants nonetheless assumed that uncertainty was present (Section 4.2.6). By explicitly visualizing uncertainty, designers can ground these assumptions in the data and more effectively support accurate reasoning about the plausible range of outcomes.

In conclusion, these findings underscore that visualization design choices strongly shape the strategies viewers adopt. Although no single strategy is universally optimal, our results demonstrate that specific visual encodings systematically encourage particular reasoning approaches, sometimes in ways unintended by designers and not aligned with normative reasoning. This highlights both the power and responsibility of visualization design: thoughtful choices can scaffold intended strategies for interpreting multiple forecasts, whereas poorly matched designs may elicit unanticipated approaches. Importantly, our results provide empirical guidance to help designers make informed decisions about the strategy-level implications of their forecast communication choices.

5.1 Future Work and Limitations

Applying the elicitation measure from Speirs-Bridge et al. [49] yielded consistent findings when participants directly placed lines on the stimuli. This direct-annotation approach differs from prior expert-elicitation studies, where experts typically report numerical values rather than marking estimates directly on a visualization. We adopted annotation because pilot testing indicated that it was easier for the general public. However, this choice introduces important considerations: asking participants to explicitly form and externalize a prediction may change how they reason about the forecasts compared to more naturalistic settings. Externalizing a judgment could either shift interpretation or, alternatively, help participants clarify otherwise imprecise intuitions. Future work should examine the benefits and limitations of prompting laypeople to make explicit visual annotations.

Participants also struggled with question 4, which asked, "How confident are you that your range, from lowest to highest, could capture the true outcome of the future event?" We do not report the results of this question due to substantial variability, suggesting that it was likely confusing or exceeded participants' knowledge. Unlike experts, who could calibrate their confidence intervals effectively [13, 16], the general public faced challenges, highlighting a potential gap in understanding that future studies could address. This question's data may be found in the Supplemental Materials.

Along similar lines, this work's scope does not include testing multiple forecast visualizations in different contexts. In the experiments we present, we prioritized rigorous experimental controls over ecological validity. For multiple forecast visualizations, further research is needed to test their use in more contexts, as the impact of individual forecasts is likely highly dependent on the specific context in which they are presented. However, this study provides foundational knowledge collected in a controlled environment, which can serve as a baseline to determine how these findings translate to more ecologically valid settings.

Our investigation was also constrained by the specific design choices tested. We only examined scenarios in which multiple forecast visualizations were co-located in one plot. The applicability of our findings to situations for which forecasts are distributed across several plots, as in the case of small multiples, remains unclear. Furthermore, the limited depth in Experiment 2 suggests the need for a more comprehensive examination of how response patterns adjust with an increase in the number of forecasts presented. Future research should pursue a more detailed analysis in this area.

6 CONCLUSION

Our study offers insights into how multiple forecast visualizations influence public interpretations of forecasts. We found that small changes in visual design and forecast selection can significantly alter viewers' interpretation strategies, demonstrating that humans do not always combine forecasts in statistically optimal ways. Our research emphasizes the importance of aligning visualizations with human reasoning by documenting interpretation strategies and the conditions under which they change.

REFERENCES

[1] Azzalini A. Azzalini. 2023. *The R package* sn*: The skew-normal and related distributions such as the skew-t and the SUN (version 2.1.1).* Università degli Studi di Padova, Italia. https://cran.r-project.org/package=sn Home page: http://azzalini.stat.unipd.it/SN/.

[2] M Bacharach. 1979. Normal Bayesian dialogues. *J. Amer. Statist. Assoc.* 74 (1979), 837–846.

[3] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods* 10, 4 (2005), 389.

[4] Alexander P Boone, Peri Gunalp, and Mary Hegarty. 2018. Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. *Journal of experimental psychology: applied* 24, 3 (2018), 275.

[5] Cynthia A. Brewer and ESRI Press. 2006. *Designing better maps: A guide for GIS users.* Taylor & Francis, Redlands, California.

[6] David V Budescu and Adrian K Rantilla. 2000. Confidence in aggregation of expert opinions. *Acta psychologica* 104, 3 (2000), 371–398.

[7] Paul-Christian Bürkner. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80, 1 (2017), 1–28. https://doi.org/10.18637/jss.v080.i01

[8] Spencer C Castro, P Samuel Quinan, Helia Hosseinpour, and Lace Padilla. 2021. Examining effort in 1D uncertainty communication using individual differences in working memory and NASA-TLX. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 411–421.

[9] Robert T Clemen and Robert L Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk analysis* 19 (1999), 187–203.

[10] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences.* routledge, New York.

[11] Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2142–2151.

[12] Stephanie Deitrick and Elizabeth A Wentz. 2015. Developing implicit uncertainty visualization methods motivated by theories in decision science. *Annals of the Association of American Geographers* 105, 3 (2015), 531–551.

[13] Ward Edwards, Ralph F Miles, and Detlof Von Winterfeldt. 2007. Advances in decision analysis. *Cambridge, New York* 1, 1 (2007), 1–638.

[14] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal, QC, Canada, 1–12.

[15] Racquel Fygenson, Lace Padilla, and Enrico Bertini. 2025. Cognitive Affordances in Visualization: Related Constructs, Design Factors, and Framework. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 10624 – 10639.

[16] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. 2005. Statistical methods for eliciting probability distributions. *Journal of the American statistical Association* 100, 470 (2005), 680–701.

[17] Miriam Greis, Aditi Joshi, Ken Singer, Albrecht Schmidt, and Tonja Machulla. 2018. Uncertainty visualization influences how humans aggregate discrepant information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal, QC, Canada, 1–12.

[18] Margaret Grounds, Susan Joslyn, and Kyoko Otsuka. 2017. Probabilistic Interval Forecasts: An Individual Differences Approach to Understanding Forecast Communication. *Advances in Meteorology* 2017 (09 2017), 1–18. https://doi.org/10.1155/2017/3932565

[19] Jake M Hofman, Daniel G Goldstein, and Jessica Hullman. 2020. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 chi conference on human factors in computing systems*. ACM, New York, NY, USA, 1–12.

[20] Jessica Hullman, Alex Kale, and Jason Hartline. 2025. Underspecified Human Decision Experiments Considered Harmful. *ACM CHI* 254, 1 (2025), 1–15.

[21] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. 2017. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 446–456.

[22] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS One* 10, 11 (2015), e0142444.

[23] Nicole Jardine, Brian D Ondov, Niklas Elmqvist, and Steven Franconeri. 2019. The perceptual proxies of visual comparison. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1012–1021.

[24] Susan Joslyn, Lou Nemec, and Sonia Savelli. 2013. The benefits and challenges of predictive interval forecasts and verification graphics for end users. *Weather, Climate, and Society* 5, 2 (2013), 133–147.

[25] Susan Joslyn and Sonia Savelli. 2021. Visualizing Uncertainty for Non-Expert End Users: The Challenge of the Deterministic Construal Error. *Frontiers in Computer Science* 2 (2021), 58. https://doi.org/10.3389/fcomp.2020.590232

[26] Alex Kale, Matthew Kay, and Jessica Hullman. 2020. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics* PP (10 2020), 1–1. https://doi.org/10.1109/TVCG.2020.3030335

[27] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 892–902.

[28] Matthew Kay. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models.* https://doi.org/10.5281/zenodo.1308151 R package version 2.3.1.

[29] Matthew Kay. 2021. *ggdist: Visualizations of Distributions and Uncertainty.* https://doi.org/10.5281/zenodo.3879620 R package version 2.4.0.

[30] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 5092–5103.

[31] John K Kruschke. 2021. Bayesian analysis reporting guidelines. *Nature human behaviour* 5, 10 (2021), 1282–1291.

[32] Dirk Leffrang and Oliver Müller. 2021. Should I Follow this Model? The Effect of Uncertainty Visualization on the Acceptance of Time Series Forecasts. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, New York, NY, USA, 20–26. https://doi.org/10.1109/TREX53765.2021.00009

[33] Le Liu, Alexander P Boone, Ian T Ruginski, Lace Padilla, Mary Hegarty, Sarah H Creem-Regehr, William B Thompson, Cem Yuksel, and Donald H House. 2016. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics* 23, 9 (2016), 2165–2178.

[34] Le Liu, Lace Padilla, Sarah H Creem-Regehr, and Donald H House. 2018. Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 882–891.

[35] Grant McKenzie, Mary Hegarty, Trevor Barrett, and Michael Goodchild. 2016. Assessing the effectiveness of different visualizations for judgments of positional uncertainty. *International Journal of Geographical Information Science* 30, 2 (2016), 221–239.

[36] Eric Newburger, Michael Correll, and Niklas Elmqvist. 2022. Fitting bell curves to data distributions using visualization. *IEEE Transactions on Visualization and Computer Graphics* 29, 12 (2022), 5372–5383.

[37] A O'hagan and Tom Leonard. 1976. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* 63, 1 (1976), 201–203.

[38] Yasmina Okan, Eva Janssen, Mirta Galesic, and Erika A Waters. 2019. Using the short graph literacy scale to predict precursors of health behavior change. *Medical Decision Making* 39, 3 (2019), 183–195.

[39] Brian D Ondov, Fumeng Yang, Matthew Kay, Niklas Elmqvist, and Steven Franconeri. 2020. Revealing perceptual proxies with adversarial examples. *IEEE transactions on visualization and computer graphics* 27, 2 (2020), 1073–1083.

[40] Lace Padilla, Racquel Fygenson, Spencer C Castro, and Enrico Bertini. 2022. Multiple forecast visualizations (mfvs): Trade-offs in trust and performance in multiple covid-19 forecast visualizations. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 12–22.

[41] Lace Padilla, Helia Hosseinpour, Racquel Fygenson, Jennifer Howell, Rumi Chunara, and Enrico Bertini. 2022. Impact of COVID-19 forecast visualizations on pandemic risk perceptions. *Scientific reports* 12, 1 (2022), 1–14.

[42] Lace Padilla, Matthew Kay, and Jessica Hullman. 2021. *Uncertainty Visualization.* Wiley StatsRef: Statistics Reference Online, Chichester, UK, 1–18. https://doi.org/10.1002/9781118445112.stat08296 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat08296

[43] Lace Padilla, P Samuel Quinan, Miriah Meyer, and Sarah H Creem-Regehr. 2017. Evaluating the impact of binning 2d scalar fields. *IEEE transactions on visualization and computer graphics* 23, 1 (2017), 431–440.

[44] Lace M Padilla, Ian T Ruginski, and Sarah H Creem-Regehr. 2017. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications* 2, 1 (2017), 1–16.

[45] R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[46] Luke F Rinne and Michèle MM Mazzocco. 2013. Inferring uncertainty from interval estimates: Effects of alpha level and numeracy. *Judgment and Decision Making* 8, 3 (2013), 330.

[47] Ian T Ruginski, Alexander P Boone, Lace M Padilla, Le Liu, Nahal Heydari, Heidi S Kramer, Mary Hegarty, William B Thompson, Donald H House, and Sarah H Creem-Regehr. 2016. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation* 16, 2 (2016), 154–172.

[48] Abhraneel Sarma, Maryam Hedayati, and Matthew Kay. 2025. More Forecasts, More (Decision) Problems: How Uncertainty Representations for Multiple Forecasts Impact Decision Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14.

[49] Andrew Speirs-Bridge, Fiona Fidler, Marissa McBride, Louisa Flander, Geoff Cumming, and Mark Burgman. 2010. Reducing overconfidence in the interval judgments of experts. *Risk Analysis: An International Journal* 30, 3 (2010), 512–523.

[50] M Stone. 1961. The opinion pool. *Annals of Mathematical Statistics* 32 (1961), 1339–1342.

[51] Xiaoqian Wang, Rob J Hyndman, Feng Li, and Yanfei Kang. 2023. Forecast combinations: An over 50-year review. *International Journal of Forecasting* 39, 4 (2023), 1518–1547.

[52] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York, NY, USA. https://ggplot2.tidyverse.org

[53] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4, 43 (2019), 1686. https://doi.org/10.21105/joss.01686

[54] Fumeng Yang, Mandi Cai, Chloe Mortenson, Hoda Fakhari, Ayse D Lokmanoglu, Jessica Hullman, Steven Franconeri, Nicholas Diakopoulos, Erik C Nisbet, and Matthew Kay. 2023. Swaying the Public? Impacts of Election Forecast Visualizations on Emotion, Trust, and Intention in the 2022 US Midterms. *IEEE Transactions on Visualization and Computer Graphics* 1, 1 (2023), 23–33.

[55] Lei Yuan, Steve Haroz, and Steven Franconeri. 2019. Perceptual proxies for extracting averages in data visualizations. *Psychonomic bulletin & review* 26 (2019), 669–676.

[56] Maximilian Zellner, Ali E Abbas, David V Budescu, and Aram Galstyan. 2021. A survey of human judgement and quantitative forecasting methods. *Royal Society open science* 8, 2 (2021), 201187.

[57] Yixuan Zhang, Yifan Sun, Lace Padilla, Sumit Barua, Enrico Bertini, and Andrea G Parker. 2021. Mapping the Landscape of COVID-19 Crisis Visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 608, 23 pages. https://doi.org/10.1145/3411764.3445381