More Forecasts, More (Decision) Problems: How Uncertainty Representations for Multiple Forecasts Impact Decision-Making

Abhraneel Sarma*
abhraneel@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Maryam Hedayati
maryam.hedayati@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Matthew Kay
mjskay@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Abstract

Users often have access to multiple forecasts regarding an event. Different forecasts incorporate different assumptions and epistemic information. A growing body of work argues against decision-making solely based on expected utility maximisation strategies in multiple forecasts scenarios, in favour of other strategies such as the maximin expected utility. In this work, we compare two different approaches for depicting epistemic uncertainty—ensembles (a direct representation of multiple forecasts) and p-boxes (a representation which only communicates the bounds of epistemic uncertainty)—in plots where individual distributions are represented as cumulative distribution plots (CDFs). We conduct three experiments to investigate the impact of the visual representation on the decision-making strategies that people adopt. Our results suggest that participants adopt conservative decision-making strategies (i.e. place greater weight on the worst-case forecast than the best-case forecast) for both p-boxes and ensembles if the set of forecasts are uniformly distributed. However, if a majority of the forecasts are clustered near one of the bounds, participants may discount the forecast which appears as a visual outlier.

CCS Concepts

• **Human-centered computing** → **Empirical studies in visualization**; **Visualization design and evaluation methods**.

Keywords

multiple forecasts, uncertainty visualization, decision-making

ACM Reference Format:

Abhraneel Sarma, Maryam Hedayati, and Matthew Kay. 2025. More Forecasts, More (Decision) Problems: How Uncertainty Representations for Multiple Forecasts Impact Decision-Making. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3706598.3713725

1 Introduction

Consider the following scenario—you are deciding whether or not to carry an umbrella before going out for the day. You receive a notification from Apple Weather on your phone that there is a high chance of rain (say 60%) in your location. Now, while you do not want to carry a large umbrella around with you all day, you'd also prefer not to get wet. In such a scenario, you could come up with utilities for each possible action and outcome, and then, taking the probability of rain into account, determine an action which will maximise your expected utility—this action would be the optimal decision. However, you also have Accuweather and The Weather Channel apps on your phone, and according to those apps, the probability of rain is 30% and 25%, respectively. You now have three different estimates for the probability of rain, and you do not have any reason to trust one forecast more than the others. Will you take your umbrella with you? How should you decide?

As probabilistic forecasts have become more commonplace and widely used for decision-making, we now often have access to more than one forecast for any particular event. Examples include COVID-19 death projections [3, 13] presidential elections [26], or even more mundane events such as the daily high temperature or the chance of rain. More importantly, the different agencies reporting these forecasts tend to use slightly different models, incorporating different assumptions or domain knowledge [3, 16, 17], which can result in somewhat varying estimates or predictions.[1]

These forecasts integrate two kinds of uncertainty: one is the quantifiable uncertainty arising due to statistical variability,[2] which we will refer to as *probabilistic uncertainty*; the other is the "uncertainty about the modeling process as a description of reality" [57] which, following Ferson and Siegrist [12], we will refer to as *incertitude*.[3] In a single forecast scenario, the only form of uncertainty is probabilistic. Thus, a decision-maker who is acting based on the Expected Utility Maximisation principle (see §2.1) would obtain the best results [46, 53]. However, with multiple forecasts, a decision-maker encounters both probabilistic uncertainty and incertitude. To use the Expected Utility Maximisation principle (or any other decision strategy) in this scenario would require the decision-maker to resolve or reconcile the two different forms of uncertainty, often through additional assumptions. For example, the decision-maker might decide to weigh the forecast with the highest chance of rain more heavily in making their decision.

Unfortunately, there is no consensus on how people *should* behave with such ambiguous information. Researchers in decision

*Corresponding Author: Abhraneel Sarma. Email: abhraneel@u.northwestern.edu

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713725

[1] For instance, during the peak of the pandemic, FiveThirtyEight maintained a dashboard of COVID-19 death projections from 10 different institutions [3], and outlined the different assumptions that are made by each of these models.

[2] There have been numerous attempts at developing taxonomies for categorising the types of uncertainty which often involve some combinations of the words *aleatory*, *epistemic* and *ontological*. Here, going by the definitions proposed by Spiegelhalter [57], by statistical variability we refer to uncertainty which is not *ontological* but overlaps with the offered definitions of *aleatory* and *epistemic* uncertainty to a certain degree

[3] For consistency, this type of uncertainty is also referred to as *epistemic* or, based on Spiegelhalter's [57] definitions, *ontological* uncertainty. While the addition of a new term(s) is admittedly only contributing to this terminology hell, we do not understand nor share the fondness of Latin terms that others in academia seemingly possess.

theory and other fields have proposed numerous strategies which may be suitable, depending on the decision-making context. One approach, known as the *principle of indifference* [55], is to not distinguish between the two types of uncertainty, and consider each forecast to be equally likely. The decision-maker can then assume a uniform prior distribution over the set of forecasts, allowing them to reduce the uncertainty from multiple distributions to a single distribution, and maximise expected utility. This approach has been the primary view adopted by recent work in visualisation and HCI [20, 32, 38]. Alternatively, a decision-maker can treat the two types of uncertainty as being qualitatively different. Instead of adopting a uniform prior distribution over the set of forecasts, this perspective proposes approaches [e.g., 7, 14, 19, 23, 29, 30, 34] which place greater emphasis on the outcome under the worst case scenario or outcomes under a wide range of possible scenarios, and better reflect a decision-maker's aversion to incertitude (e.g., the maximin criterion for expected utility).

Just as there are different ways of visualising probabilistic uncertainty (e.g., confidence intervals, probability density plots, dotplots, cumulative distribution plots etc.), there are also different ways of visually depicting incertitude—e.g., *ensembles* and *probability boxes* (or p-boxes). Ensembles are commonly used to visualise multiple forecasts, and they faithfully depict the distributional information from each forecast. However, prior work has argued that ensembles may be misinterpreted, with the viewer perceiving more frequently occurring distributions as more likely [22, 49]. As an alternative, Ferson and Siegrist [12] propose p-boxes to communicate both probabilistic uncertainty and incertitude, without, at least theoretically, conflating the two. P-boxes (see Figure 1) are specified by the left and right bounds of the cumulative distribution functions of a set of uncertainty distributions (forecasts). While we cannot resolve theoretical disagreements on how to treat multiple forecasts, we hypothesise that **different visual representations of incertitude will lead to behaviours that are better aligned with different decision-making strategies**.

Neither ensembles nor p-boxes have been empirically evaluated on their impact on decision-making under multiple forecasts. We aim to measure that alignment so that decision-makers can adopt visualisations of incertitude that better match their decision criteria. Specifically, the **objective of this paper is to evaluate the impact of using different visualisations of incertitude on users' decision-making strategies** in an incentivised decision-making task, for a specific representation of probabilistic uncertainty—cumulative distribution plots (CDFs). We conduct three pre-registered experiments. In Experiment 1, we compare decision-making when participants are presented with either p-boxes or ensembles of approximately uniformly distributed forecasts. The results suggest that both p-boxes and ensembles can lead to participants adopting uncertainty-averse decision-making strategies where the worst-case scenario is weighted more heavily. In Experiment 2, we examine how the distribution of the forecasts in an ensemble representation impact decision-making and present participants with ensembles of forecasts which are either skewed left or skewed right. We find that participants' decision-making strategies are predominantly based on the main cluster of forecasts but were still somewhat uncertainty-averse on average. We conducted Experiment 3 as a robustness check to assess the impact of a difference in phrasing between the two conditions (p-boxes and ensembles) in Experiment 1, and found no evidence to suggest an effect. Taken together, our results suggest that, unlike p-boxes, in the case of ensemble representations, the decision-making strategies participants adopt will likely vary based on how the forecasts are distributed. Based on our results, we outline a set of design recommendations for representing multiple forecast distributions.

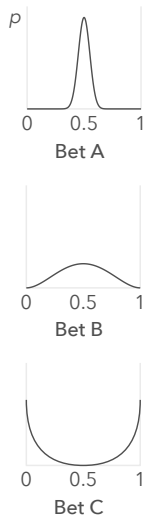
2 Background

Our work draws upon: (1) prior work about various decision-making strategies under incertitude in fields such as decision theory and risk analysis; and (2) uncertainty visualisation approaches.

2.1 Perspectives from Decision Theory and Risk Analysis

In decision theory, researchers typically distinguish between uncertainty which can be quantified or statistically characterised (which is typically referred to as risk in economics, and as *aleatory uncertainty* in statistics) from uncertainty which is unquantifiable [57] (which is referred to as *ambiguity* [10], Knightian uncertainty [28], or even epistemic uncertainty [43]). In risk analysis, "situation(s) where the decision-makers do not know or cannot agree on a single probability density function of the outcomes" are referred to as *deep uncertainty* [34]. Using this lens, the uncertainty in a single forecast, which is often described using a probability distribution function, can be considered *probabilistic uncertainty*; the uncertainty due to lack of perfect knowledge about the data generating process, which results in multiple forecasts each encoding different assumptions about the data generating process, can be considered *incertitude*. A consequence of incertitude is that probability for a particular outcome is either unknown or cannot be precisely stated.

Decision-making in scenarios with both forms of uncertainty requires the decision-maker to make certain assumptions to resolve the two forms of uncertainty. A common assumption is the *principle of indifference* [55], which suggests that in the absence of any other information, the decision-maker should not distinguish between the two types of uncertainty and consider each forecast to be equally likely. The decision-maker can then assume a uniform prior over the set of forecasts, and consider the average of all the distributions when making a decision. Recent work in visualisation and HCI [20, 32, 38] have primarily adopted this perspective. However, there are three arguments against this principle: (1) it fails to distinguish between the two types of uncertainties, which many argue are qualitatively different; (2) it fails to take into account for people's aversion to uncertainty; and (3) it is sensitive to how many forecasts are available to the decision-maker.



Gardenfors and Sahlin [21] argue that people treat the two forms of uncertainty differently. They provide an example where a decision-maker is asked to bet, at equal odds, on the outcome of three events—the forecasts for each event estimates the expected probability to be the same

(0.5), but the forecasts themselves have different degrees of uncertainty (see adjacent figure). Gardenfors and Sahlin argue that while a decision-maker is likely to take the first bet, they are unlikely to do so for the second and third bets as these forecasts are *less reliable*. The Ellsberg paradox [10] demonstrates that in a game of chance where the underlying probabilities are unknown, participants employ decision rules which are incompatible with expected utility maximisation. One explanation of people's behaviour in both of these examples is their aversion to incertitude [19, 41].

Decision theorists have proposed alternative strategies which allow the decision-maker to transform the problem of incertitude—decision-making under multiple uncertainties—into a problem of partial certainty [e.g., 19, 21, 23, 27, 33]. One such strategy is the *maximin criterion for expected utility* (MMEU), which recommends the action with the largest minimal expected utility [19, 21]. Another is the *optimism-pessimism rule*, which trades off a little bit of certainty for potentially higher utilities by asking the decision-maker to weigh both the best and worst possible outcomes for each alternative [23, 41] according to an optimism parameter. We discuss these strategies in more detail in §3.

In risk analysis, Lempert and Collins [30] argue that while expected utility maximisation has been widely adopted for many problems in this domain, it only yields the best answer if the "uncertainty is well characterised." However, for many decision problems uncertainty information may be imprecise, ambiguous, or even absent. Researchers seek to identify robust decision-making strategies which are predicated on the assumption of precaution and taking into consideration performance in the worst-case scenario [30, 34]. Many of these strategies have their roots in decision theory [30]. For instance, the Limited Degree of Confidence strategy [7, 14, 29], which maximises a weighted average of the expected utility and the utility in the worst-case scenario is similar to, and can be expressed in terms of the optimism-pessimism rule.

2.2 Visualising multiple forecasts

Probabilistic information is frequently represented using visual representations such as text [e.g., 11, 18, 65, 66], confidence intervals [e.g., 2, 11, 25, 52, 66], probability density plots [e.g., 11, 25, 52], dotplots [e.g., 11, 25, 66] and cumulative distribution plots [e.g., 11, 24], and some of these uncertainty visualisations can improve decision quality under uncertainty [e.g., 11, 25, 37, 54]. To accurately visualise multiple forecasts, we need to communicate both probabilistic uncertainty in the prediction of a single forecast, as well as incertitude due to presence of multiple forecasts. Forecasts from multiple models have been commonly depicted using ensembles of representations for singular probability distributions (Figure 1) [9, 32, 39, 48]. However, when used to represent forecasts from multiple models, ensembles may be misconstrued as providing probability or frequency information—readers may consider a greater frequency of similar curves to represent forecasts which are more probable. While this might be the desired interpretation in some cases [e.g., 9, 39, 47], in others, such as the scenarios being considered in this study, that is not the case.

To avoid misinterpreting more frequent forecasts as being more probable, we consider other graphical representations which would allow us to **vary how incertitude is visualised for the same probabilistic uncertainty representation**. Ferson and Siegrist [12] propose probability bounds analysis as *a calculus for imprecise probabilities*, and propose *probability boxes* (p-boxes) as an alternative approach to visualising multiple forecasts consistent with the notions of probability bounds analysis. A p-box is specified by the "left and right bounds on the cumulative probability distribution function of a quantity" (see Figure 1) and communicates variability (i.e. probabilistic uncertainty) as the *slant* of any curve within the bounds; it communicates incertitude through the width between the left and right edges of the box [12].

Bounds on cumulative distribution plots (CDFs) are formally consistent with the notions of worst-case (and best-case analysis)—they represent well-formed objects that can be manipulated by the calculus of imprecise probabilities [12]. For example, a p-box constructed from a set of CDFs represents a set of probability distributions—any (monotonically increasing) curve which can be contained within the bounds. Bounds constructed using other uncertainty representations such as probability density plots (PDFs) will not capture the same set of distributions as bounds on CDFs (see Figure 2), and therefore do not correspond to well-formed objects in the calculus for imprecise probabilities—the (visual) operator of finding valid members of a p-box (any monotonically increasing curve) is not valid for bounds on PDFs. Additionally, while the construction of a p-box from a CDF is straightforward, and results in an elegant, unified representation of both probabilistic uncertainty and incertitude, bounds constructed from PDFs can result in idiosyncratic shapes (e.g., the p-box constructed from the distributions in Figure 2 results in a bimodal shape).

For the tasks in our experiment (estimating tail probabilities), prior work has found CDFs to be highly effective [24]. CDFs allow a viewer to directly read tail probabilities (e.g., $\Pr(T < 0)$) on the y-axis, whereas estimating tail probabilities from other representations, such as PDFs, require the viewer to make less accurate area comparisons [36, 59, 60]. CDFs have also been found to be effective in decision-making tasks similar to the one used in the current study [11]. Other representations such as intervals, while commonly used, provide viewers with limited distributional information compared to CDFs and PDFs. Finally, consonance curves can also be used to visualise bounds on a set of distributions in a consistent manner [49], but we considered this representation to be more complex than CDFs, and omitted it from our current study.

In theory, by suppressing information regarding the frequency of each individual forecast distribution, p-boxes, according to the

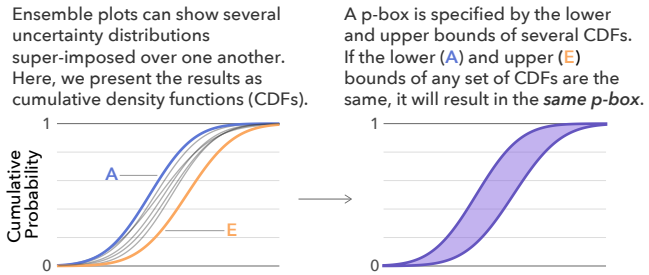


Figure 1: The two representations considered in the study, ensembles of CDFs and p-boxes.

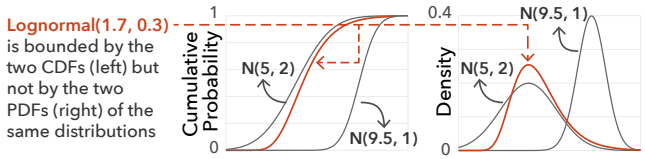


Figure 2: Bounds constructed using Cumulative Density Plots (CDFs) and Probability Density Plots (PDFs) are not equivalent, as certain distributions which are within a p-box constructed from CDFs will not be within the bounds of PDFs.

expressiveness principle [36], should not allow a reader to draw probabilistic conclusions about the forecasts, thereby potentially overcoming the limitations of representations such as ensembles. However, whether viewers of p-boxes interpret the visualisation as intended has not been empirically evaluated. Moreover, whether viewers of ensemble CDF plots indeed misinterpret the presented information (i.e., assume an incorrect probabilistic interpretation) as claimed has also not been empirically evaluated. We aim to address this gap in the current work.

3 Decision-Making Under Single and Multiple Forecasts

We describe the decision-making strategies for both single and multiple forecast scenarios using the decision task that we presented to participants.

Single forecast. We adapt the following hypothetical scenario from Padilla et al. [40]:

> Assume that you work at the Red Cross, and your job is to manage resources for farms in Peru. In previous years, alpacas have died in Peru from cold temperatures. Alpacas can typically withstand the cold unless the temperature drops below 0°C (32°F). You are in charge of the Red Cross's blanket budget, and it is your job to issue blankets to the alpacas when temperatures fall below 32°F, which will help them withstand the cold. You have a budget for 18 days of $18,000. Purchasing and delivering blankets to farmers costs $1,000 (per night). If you fail to issue blankets to the farmers and the temperature drops below 32°F, it will cost $5,000 from your budget. You are shown a night-time temperature forecast distribution. Based on this forecast, you have to decide whether to issue blankets to the alpacas.

The payoff matrix for the decision problem described above can be represented using the following table:

	$s_1 : T \leq 0$°C (32°F)	$s_2 : T > 0$°C (32°F)
a_1 : send	-1000	-1000
a_2 : ¬send	-5000	0

Here, $S = \{s_1, s_2\}$ represents the two possible states of nature, and $A = \{a_1, a_2\}$ represents the two possible actions that are available to the decision maker. The incentives correspond to a utility function $U(a_i, s_j) = u_{ij}$ defined over S and A. Finally, the temperature forecast provided is a single probability distribution $\mathrm{P} : S \rightarrow [0, 1]$. In other words, P gives the probability of each state s_i occurring. Thus, the expected utility of any action a_i is $\mathbb{E}[U(a_i)] = \sum_j U(a_i, s_j) \cdot \mathrm{P}(s_j)$. Based on this incentive scheme, an action a_1 (to send aid to the alpacas) is preferred if and only if:

$$\mathbb{E}(U(a_1)) \geq \mathbb{E}(U(a_2))$$
$$u_{11}P(s_1) + u_{12}P(s_2) \geq u_{21}P(s_1) + u_{22}P(s_2)$$
Let, $P(s_1) = p$; Since, $P(s_2) = 1 - P(s_1)$
$$u_{11}p + u_{12}(1-p) \geq u_{21}p + u_{22}(1-p) \quad (u_{22} = 0)$$
as $u_{11} = u_{12}$
$$u_1 \geq u_2 p \quad (1)$$
$$-1000 \geq -5000p$$
$$p \geq 0.2 \quad (2)$$

Thus, the utility-optimal decision in this scenario is to send blankets if the probability of freezing is greater than or equal to 0.2. We refer to $p = 0.2$ as the optimal *crossover point*, as it represents the probability at which a decision maker should not have a preference over the two actions. Also note that going forward, wherever possible, we will be adopting simplified notation for utilities and probabilities according to (1).

Multiple forecasts. Now consider the analogous scenario where the decision maker is instead presented with multiple forecasts:

> You will be presented with forecasts from seven different agencies, all of which are considered to be reliable. Based on the forecasts, you will be asked whether you will issue blankets to the alpacas.

In this scenario, we still have the same set of possible states (S), the same set of possible actions (A) and the same utilities for each outcome. However, instead of a single probability distribution describing the uncertain state of nature, we now have a set of probability distributions, $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$. As a result, we cannot determine the expected utility of an action $\mathbb{E}[U(a_i)]$ unless we make assumptions regarding the probability of each forecast in $\mathcal{P}$, or determine rules which can help reduce the uncertainty that is faced by the decision maker. Readers might contend whether or not these instructions induce a (uniform) prior on the participants. However, this is espoused in the objectives of this paper—we expect participants with different philosophies to interpret the information differently, and we investigate whether the method of visualising the forecasts impacts which philosophy participants adopt. Below, we outline rules (which stem from different philosophical interpretations of incertitude) that have been put forth in the decision-theoretic literature (discussed in §2.1).

The Principle of Indifference: In the absence of any further information about the quality of the forecasts, one approach for the decision-maker would be to consider each forecast as equally likely and assume a uniform distribution over the set of forecasts in $\mathcal{P}$. Mathematically, the utility-optimal decision, based on this decision rule, can be represented as choosing the action which maximises $\mathbb{E}[U(a_i)] = \frac{1}{n} \sum_k [u_{i1} \cdot p_k + u_{i2} \cdot (1 - p_k)]$ where p_k is the probability of freezing for a forecast $P_k \in \mathcal{P}$ and $n = |\mathcal{P}|$. However, critics of this approach have argued that it fails to adequately account for incertitude and that assuming a uniform prior over the set of forecasts is a relatively strong assumption—if there is no reason

to think that one forecast is more probable than another, then it may be arbitrary to conclude they are all *equally* probable [21, 41]. These critics argue that other decision criteria may imply it is not irrational to be more uncertainty-averse than what is suggested by this approach [19, 21, 41].

Maximin Criterion for Expected Utility (MMEU): To incorporate potentially uncertainty-averse behaviour, decision theorists have identified pessimistic decision rules such as the maximin criterion for expected utilities. This rule involves selecting the action which maximises the minimal expected utility: for each possible action a_i, the decision maker calculates the expected utility under the worst possible outcome, *i.e.* $\mathbb{E}_{min}[U(a_i)] = \min_k\{u_{i1} \cdot p_k + u_{i2} \cdot (1 - p_k)\}$; they then select the action which leads to the best worst expected utility: $\arg\max_i \mathbb{E}_{min}[U(a_i)]$. By focusing on the worst possible outcome for each action, MMEU provides a partial certainty regarding what the outcome will be (i.e., the decision-maker will, at the very least, obtain the utility associated with the worst possible outcome) [19, 41].

Maximax Criterion for Expected Utility: By the same logic as the MMEU, but instead focusing on the best possible outcome, one can also define the maximax criterion: for each possible action a_i, calculate the utility for the best possible outcome and select the action which leads to the best best expected utility, *i.e.* $\arg\max_i\{\mathbb{E}_{max}[U(a_i)]\}$. While this is not a decision rule that is typically adopted in practice, the argument is that it is no less rational than the MMEU [41].

The Optimism-Pessimism rule: Also known as the Hurwicz rule, this is a generalisation of the maximin and maximax rules. The decision maker considers the expected utilities under both the worst-case and best-case scenarios, weighted according to an optimism parameter γ. In other words, the expected utility of any action a_i is: $\mathbb{E}_\gamma[U(a_i)] = \gamma\mathbb{E}_{max}[U(a_i)] + (1-\gamma)\mathbb{E}_{min}[U(a_i)]$; they then select the action which maximises $\mathbb{E}_\gamma[U(a_i)]$.

As the optimism-pessimism rule is a generalisation of the other rules, we can use it to describe the decision-making strategies that are adopted by participants. Thus, 2, which describes when the decision to select action a_1 is preferred, can be extended to the multiple forecasts scenario as:

$$\mathbb{E}_\gamma[U(a_1)] \geq \mathbb{E}_\gamma[U(a_2)]$$

γ: optimism index

$$u_1 \geq \gamma \cdot \max_k\{u_2 \cdot p_k\} + (1-\gamma) \cdot \min_k\{u_2 \cdot p_k\}$$

p^+: upper bound of the forecasts corresponding to $\max\{u_2 \cdot p_k\}$

p_-: lower bound of the forecasts corresponding to $\min\{u_2 \cdot p_k\}$

$$u_1 \geq [\gamma \cdot (u_2 \cdot p^+) + (1-\gamma) \cdot (u_2 \cdot p_-)]$$

$$u_1 \geq u_2[\gamma p^+ + (1-\gamma)p_-]$$

$$\gamma p^+ + (1-\gamma)p_- \geq 0.2 \quad (3)$$

We use the calculations in 2 and 3 to derive a regression model to analyse the responses of participants in our experiments as a decision problem in §4.2.

4 Experiment and Analysis

We conduct three preregistered experiments to investigate our research questions. In **Experiment 1**, we compare two representations of multiple forecasts—p-boxes and ensembles—to determine whether these representations have an impact on the decision-making strategies that participants adopt (see `preregistration`). In **Experiment 2**, we examine whether the distribution of individual forecasts within an ensemble have an impact on participants' decision-making strategies by comparing two different distributions of ensembles—ensembles skewed left and ensembles skewed right(see `preregistration`). We use a mixed factorial design for both experiments.

In addition, we accidentally used the phrase "equally reliable" instead of "reliable" to describe the different forecasters in the participant instructions for the ensembles condition of Experiment 1 (this was corrected in Experiment 2). This phrasing may have had an unintended effect of inducing a uniform prior on the participants. Due to this mistake, we conducted Experiment 3 as a robustness check to ensure that our results were not impacted by the differences in instructions (see `preregistration`). To preview the results of this robustness check: we do not believe our results were impacted by this mistake (see §5.3).

4.1 Experimental Materials

In Experiment 1 and Experiment 2, there are three experimental variables: (1) the type of uncertainty representation; (2) the number of forecasts shown (*one* or *many*); and (3) the forecast itself. Both experiments used a mixed-factorial design. The number of forecasts, and the forecasts themselves, were varied within-subjects; representation was varied between subjects. In each experiment, participants were asked to perform a series of incentivised decision making tasks (trials) with the same alpaca scenario described in §3. These tasks were divided into two blocks.

Representations: In both experiments, the single forecasts in the first block of trials were represented as cumulative distribution functions (CDFs). In Experiment 1, the uncertainty representations used were p-boxes and (uniformly distributed) ensembles (see Figure 3B), varied between-subjects. In Experiment 2, the uncertainty representations were left-skewed and right-skewed ensembles (see Figure 3C), varied between-subjects.

Procedure: The first block consists of 18 trials. In each trial participants were presented with a single forecast in the form of a Normal distribution with a standard deviation of 2.08 (see Figure 3). We varied the means of the Normal distribution from 29.5°F to 38°F in intervals of 0.5 to generate different forecasts. The probability of freezing corresponding to these forecasts are: $\Pr(T \leq 32°\text{F}) = \{0.89, 0.84, 0.74, 0.68, 0.59, 0.5, 0.41, 0.30, 0.24, 0.17, 0.10, 0.07, 0.06, 0.03, 0.02, 0.009, 0.005, 0.003\}$. The second block consists of 12 trials. In each trial, participants were presented with a multiple forecast visualisation. In Experiment 1, the multiple forecasts were subsets of seven consecutive elements of the set of forecasts used for the first block. We chose these distributions so that we would be able to estimate a crossover point even if participants were adopting the extreme pessimistic (i.e. $\gamma = 0$) or optimistic (i.e. $\gamma = 1$) strategies. For Experiment 2, we used the same bound for each ensemble forecast as Experiment 1; we then randomly sampled a set of forecast

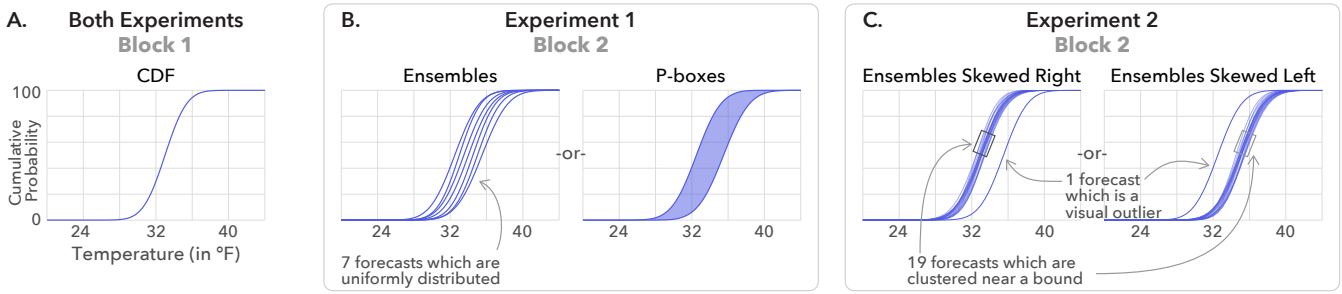


Figure 3: Illustration of the stimuli used in the two experiments. Both experiments consisted of two blocks. The first block in both experiments required participants to perform the binary decision making task based on a single forecast, visualised as CDFs. In block 2 of experiment 1, participants were presented with either p-boxes or ensembles. In block 2 of experiment 2, participants were presented with either ensembles skewed right or skewed left. The actual stimuli used in the experiment can be found in `supplement ▶ stimuli`.

distributions whose means were at most 1°F from the mean of the pertinent bound (i.e. the lower bound in the ensembles skewed right condition and the upper bound in the ensembles skewed left condition).

After each trial in the first block, we simulate a *state of nature* based on the forecast and provide participants with feedback on how well they performed. This also impacts the final payout they receive for their participation in the form of bonuses. In the second block of trials, participants were not provided with any feedback. Moreover, their performance was not evaluated (i.e., the second block of trials do not impact the payout that participants receive), but they were not informed of this, and thus were led to believe that their performance in the second block would be evaluated as well. This is because evaluating multiple forecasts would either require us to assign a probability distribution to the set of forecasts (which would require us to reify one of the decisions rules described in §3 as being correct in this experiment).

The two blocks in our experiment are necessary for obtaining precise estimates for the intercept (α), slope (β) and the optimism (γ) parameters in our linear-in-logit model (§4.2). Participants' responses in the first block of trials (where only single forecasts are shown) are used to estimate the intercept (α) and slope (β) parameters, as shown in `line 8` ; the responses in the second block of trials are used to estimate the optimism parameter. We suspect that without the two distinct blocks of trials, our model might run into identifiability issues in distinguishing between the parameters. We validate whether our model can recover these parameters (see §4.3) After completing all trials, participants answered one multiple choice question which asked them to indicate which forecast (lower bound, median, or upper bound) they primarily used for making their decisions. They were also asked to describe their strategy for performing the task using an open text field.

Tutorials and Training. In both experiments, at the onset, we provide instructions to participants on how to correctly interpret a CDF plot. This is followed by participants completing two training trials, before the actual test trials. In the training session, participants were shown the temperature forecast $T \sim Normal(67, 2)$ and were asked to report the probability of the temperature falling below 66°F and 68°F respectively. Participants were then provided feedback on whether they answered the question correctly or not. The training trials also served as a comprehension check—we excluded participants who failed to answer at least one of the questions correctly, as per our preregistration. This helped us ensure that participants were correctly reading the CDF plots, which was essential for the validity of our study [51].

Participants: We recruited all participants from Prolific. Our experiment was only eligible to participants who were fluent in English, lived in the U.S. (due to the usage of the imperial system in our stimuli), and on desktop devices. For both experiments 1 and 2, as per our pre-registrations, we aimed to recruit 150 participants (75 participants in each condition). After excluding participants who failed to meet our criteria, we had 156 participants (76 in the p-box condition and 80 in the ensembles condition) for Experiment 1, and 142 participants (71 in each condition) for Experiment 2. The median completion time was approximately 13 mins for Experiment 1 and 11.5 mins for Experiment 2, corresponding to an average wage of $14/hr and $15.5/hr, excluding bonuses.

4.2 Model specification

When asked to make decisions and judgements based on probability, participants in such studies often provide imperfect responses due to "some distortion of judgement or misperception of probabilities" [65]. According to Zhang and Maloney [67], a linear-in-log-odds (llo) model can be a good fit to account for distortions or misperceptions in probability judgements. We therefore translate equations (2) and (3) into log-odds and apply a linear transformation to derive a model of participants' decisions:

$$\text{From (2):} \quad \text{logit}(p) \geq \text{logit}(0.2)$$
$$\alpha + \beta \cdot [\text{logit}(p) - \text{logit}(0.2)] \geq 0 \quad (4)$$

$$\text{From (3):} \quad \text{logit}(\gamma p^{+} + (1-\gamma)p_{-}) \geq \text{logit}(0.2)$$
$$\text{logit}(\gamma p^{+} + (1-\gamma)p_{-}) - \text{logit}(0.2) \geq 0$$
$$\alpha + \beta \cdot [\text{logit}(\gamma p^{+} + (1-\gamma)p_{-}) - \text{logit}(0.2)] \geq 0 \quad (5)$$

We use the model formula derived above in (4) and (5) to implement a Bayesian Hierarchical linear-in-log-odds regression model. The full model formula can be specified as follows:

`line 1` $$decision \sim \text{Binomial}(p_{\text{SEND}})$$

intercept slope probability of freezing

`line 2` $$\text{logit}(p_{\text{SEND}}) = \begin{cases} \alpha_i + \beta_i \cdot [\text{logit}(p) - \text{logit}(0.2)] & \text{for block 1} \\ \alpha_i + \beta_i \cdot [\text{logit}(\gamma_i p^+ + (1-\gamma_i)p_-) - \text{logit}(0.2)] & \text{otherwise} \end{cases}$$

p^+: upper bound probability of freezing; p_-: lower bound probability of freezing

`line 3` $$\alpha_i = \alpha + \delta_{\alpha,i}$$

`line 4` $$\beta_i = \beta + \delta_{\beta,i}$$

`line 5` $$\begin{bmatrix} \delta_{\alpha,i} \\ \delta_{\beta,i} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma\right)$$

`line 6` $$\text{logit}(\gamma_i) = \omega_r + \delta_{\omega,i}$$

`line 7` $$\delta_{\omega,i} \sim \text{Normal}(0, \sigma)$$

`line 8` $$r \in \{1...R\} \quad (R \textit{ representations})$$

`line 9` $$i \in \{1...N\} \quad (N \textit{ participants})$$

Line 1: The binary decision made by a participant in each trial is modeled as a Binomial distribution with probability p_{SEND}, where 1 represents the decision to send blankets and 0 represents the decision to not send blankets.

Line 2: In the single forecast scenario, we assume that participants' decisions would be a function of the temperature forecast distribution shown, which can be represented using the mean of the distribution. In the multiple-forecast scenario, we assume that participants would be making their decision based on some distribution contained within the bounds of the set of forecasts; we identify this distribution using: $\gamma p^+ + (1-\gamma)p_-$, where $\gamma \in [0, 1]$ represents an optimism index parameter, and p^+ and p_- represent the upper and lower bounds of the probability of freezing in the set of forecasts.

Lines 3-5: We expect the intercept ($\alpha_{[i]}$) and the slope ($\beta_{[i]}$) parameters to vary between participants, as different participants will likely have different decision-making capacities. α and β are the slope parameters for the *average* participant ($\delta_{\alpha,i} = 0$; $\delta_{\beta,i} = 0$), whereas $\delta_{\alpha,i}$ and $\delta_{\beta,i}$ capture differences between each participants' intercepts and slopes compared to the average participant, as random effects.

Line 6: The optimism factor, γ, may also vary for different representations as well as for different participants. Because it is bounded ($[0, 1]$), we use a logit transformation.

Priors: As can be seen from Figure 4, perfectly unbiased responses would yield values of $\alpha_{[i]} = 0$ and $\beta_{[i]} = 1$. We thus use priors centered on these values but which also permit the possibility of significant distortion: $\alpha \sim \text{Normal}(0, 1)$ and $\beta \sim \text{Normal}(1, 1)$. We use zero-centered priors for the random effects parameters ($\delta_{\alpha,i}$ and $\delta_{\beta,i}$). We assume γ to be centered at 0.5. The prior on the average γ is determined by ω_r in logit space (`line 6`), where 0.5 corresponds to 0, so we use the prior $\omega_r \sim \text{Normal}(0, 1)$, which is centered at 0.5 in the inverse-logit space but permits values of γ very close to zero or one.

Implementation: We implemented these models in `R 4.4.0` [44] and `CmdStanR 0.8.0` [15]. The model ran four chains with 4,000 warmup samples and 4,000 post-warmup samples each, thinned by 4 for a final total sample size of 4,000. We assessed convergence using the Gelman-Rubin diagnostic ($\hat{R} = 1.00$ for all population-level parameters, correlations and standard deviations) and the (bulk and tail) effective sample sizes ($\text{ESS}_{min} \approx 3,000$).

4.3 Model Validation

Our Bayesian model is admittedly complex. Given this complexity, it is important to make sure that the model is: (1) calibrated, and (2) is able to estimate a posterior distribution which is close approximation of the observed data. We use simulation-based calibration and posterior retrodictive checks respectively to validate our Bayesian model.

Simulation-Based Calibration (SBC) is a procedure which "identifies inaccurate computation and inconsistencies in model implementations." [62]. In our model (§4.2), the primary source of complexity stems from the need to accurately estimate, and distinguish between the values of, the α, β and γ parameters in 3-5. For instance, consider the scenario where $\alpha < 0$ and $\gamma \in (0, 0.5)$—parameter values which are plausible given our prior distributions. These values would mean that the average participant both perceives the optimal crossover point to be greater than 0.2 *and* weighs the lower bound of the forecasts more. The SBC procedure allows us to validate that our implemented model is able to recover the underlying structure of this data-generating process and correctly estimate the parameters in all scenarios permitted by our priors. Using SBC, we repeatedly simulate parameters from the prior distribution, then simulate datasets using the simulated parameter values. For each simulated dataset, we fit our model, obtaining draws (say M) from the posterior. We calculate the rank of the simulated parameter values with respect to the M posterior draws. By construction, if a Bayesian model is calibrated, the rank statistics of simulated parameter values should be uniform [8, 35, 62]. Our SBC checks do not reveal any potential issues, suggesting our model is able to accurately recover the α, β, and γ parameters (see `supplement ▶ R ▶ 05-validation.Rmd`).

Posterior Predictive/Retrodictive Checks While SBC can highlight potential issues with the model implementation, it does not tell us if our model is a good fit for the observed data. We use posterior retrodictive checks—using the posterior predictive distribution from our estimated model, we retrodict existing participant responses

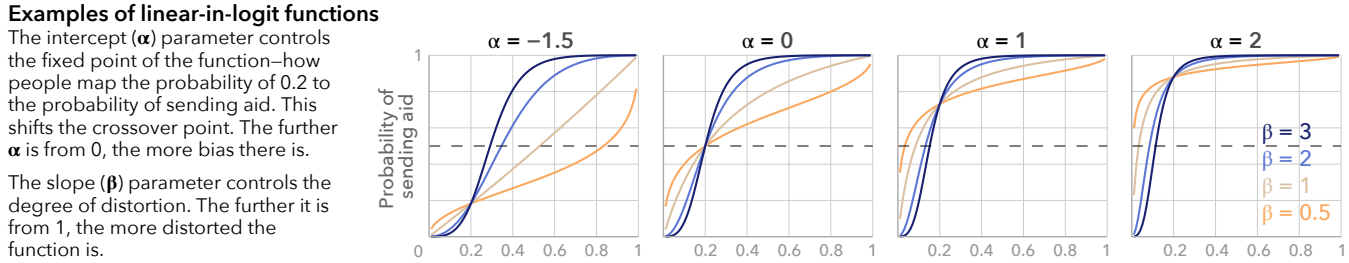


Figure 4: Examples of different possible linear-in-logit models for different values of the intercept (α) and slope (β) parameters. Figure adapted from Yang et al. [65].

[4] and compare them to the participant response averages. Our posterior retrodictive checks do not show any signs of consistent model bias and indicate a good model fit (see `supplement ▶ R ▶ 05-validation.Rmd`).

4.4 Qualitative Analysis

As per our preregistration, we conducted an exploratory qualitative analysis of participants' self-reported strategies. We primarily used their response to the open-text question, as we found participants' responses to the multiple-choice question (where they were asked to indicate which forecast they primarily used for making their decisions) were sometimes inconsistent with their textual explanation and we were interested in the strategy they used rather than only which line they looked at.[4] We used participants' multiple-choice responses only in the cases where the response helped resolve some confusion in our interpretation of their response.[5]

We used a hybrid coding approach—using our knowledge of possible strategies as well as participants' responses—to develop a set of codes. We coded participants' primary and supplementary strategies, focusing on the aspect(s) of the visualisation (usually which distribution(s)) they used to make their decision. We also noted the decision point participants' used—the optimal decision point would be 0.2 (§3), but participants listed a wide range of decision points. If participants listed a range of values as their decision point (e.g. "Once the risk got above 25–30% I felt it was worth buying the blankets"), we used the lower bound (0.25 in this example). Some people did not explicitly state the crossover point for their decisions, but we were able to infer a lower bound based on an example they provided (e.g., "I look at the graph above and see that at most you have a 30% need for blankets [...] so I wouldn't send blankets on this one", which suggested that their crossover point was at least 0.3).

We also noted whether each participant had any misunderstandings about the task, and whether they expressed a risk-seeking or risk-averse bias, or explicitly mentioned wanting to do well on the task due to their empathy for the alpacas. We discuss these additional codes in the discussion section. The second author acted as the initial coder and performed open coding to generate initial codes. Both coders discussed the codebook and reached an initial consensus, then both independently coded all responses. All disagreements were discussed until a consensus set of labels was reached. The full codebook is available in `supplement ▶ qualitative-analysis ▶ qualitative-analysis.xlsx`.

4.5 Experiment Details For Experiment 3

The primary difference between Experiment 3 and the previous experiments was that instead of manipulating the uncertainty representation, we used the same representation (ensembles) and only manipulated the phrasing used to describe the forecasters (*reliable* or *equally reliable*) between subjects. We kept every other detail regarding the design of the experiment consistent. For our analysis, the only change was in `line 8` where instead of different representations, we had different phrasings as the between-subjects manipulation. As per our preregistration, we recruited 99 participants (49 in the *reliable* condition and 50 in the *equally reliable* condition). The median completion time was approximately 9 mins, corresponding to an average wage of $20/hr, excluding bonuses.

5 Results

5.1 Experiment 1

The results of Experiment 1 (Figure 5C-D) show that the values of γ are 0.23 (95% CI: [0.16, 0.29]) and 0.26 (95% CI: [0.18, 0.32]) for p-boxes and ensembles respectively, suggesting a small, but practically negligible difference (-0.02; 95% CI: [-0.12, 0.07])) in the optimism parameter between the two visualisation conditions. This means that both p-boxes and ensembles led to similarly cautious, uncertainty-averse, decision-making, which weighs the lower bound (worst-case forecast) more than the upper bound (best-case forecast). We estimate the *decision-making reference distribution*—the approximate distribution that participants are using to make their decisions if they are employing some form of weighting of the forecasts. Figure 5D shows that this decision-making reference distribution is virtually indistinguishable between p-boxes and ensembles.

Additionally, the estimates of the α (-1.62; 95% CI: [-1.96, -1.30]) and β (2.86; 95% CI: [2.55, 3.20]) parameters suggests that participants exhibit some degree of bias in their responses (Figure 5A). The estimated crossover point—the point at which a participant should

[4] e.g., one participant listed the lower bound as their multiple choice response but stated that they were "generally looking at [the median/average curve] as most of the graphs overlapped there", and that "for this specific graph, [they were] looking at [lower bound] for worst case scenario". We therefore coded their primary strategy as using the majority of the lines.

[5] e.g., one participant said "if the graph showed anywhere higher than 55%-ish then [they] decided to send blankets just to be safe", and their multiple choice response was the lower bound curve, confirming that they were referring to the lower bound when saying "anywhere higher".

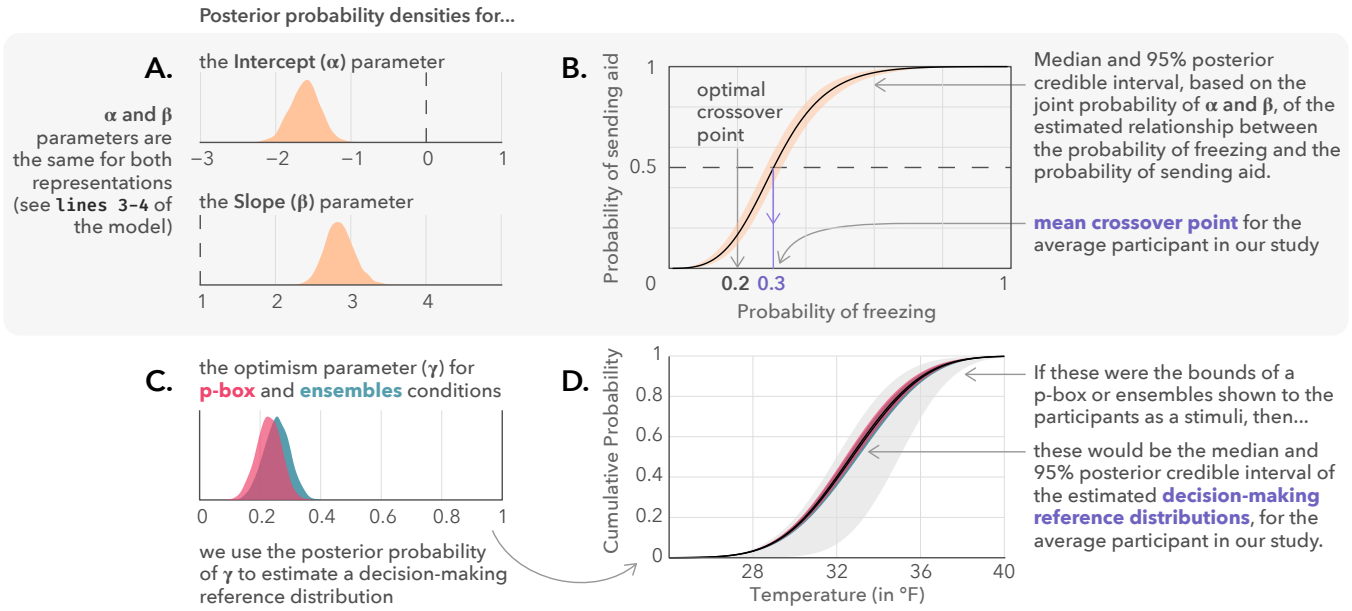


Figure 5: The main result of Experiment 1. We present the posterior probability estimates of the α and β parameters (A), and the γ parameter for the two conditions (C); the median and 95% credible interval of the estimated linear relationship based on joint distribution of α and β (B), and the estimated decision-making reference distribution (D).

have no preference between the two choices—is when the forecasted probability of freezing is 0.31 (95% CI: [0.28, 0.33]), which is greater than the optimal crossover point of 0.2 (Figure 5B). This might indicate that participants may be misidentifying the optimal crossover point. These parameter estimates are consistent with Yang et al. [65], where participants exhibit a bias in the same direction in both conditions, and Padilla et al. [40], where participants exhibit a bias in the same direction in 2/4 conditions. The analysis for Experiment 1 can be found in `supplement ▶ R ▶ 04-analysis.Rmd`.

In our qualitative analysis of participants' self-reported strategies in the p-box condition, 32 participants reported primarily using the lower bound forecast to make their decision, as evidenced by statements like: "I used [the lower bound forecast] mostly because it's better to be safe than sorry when dealing with alpacas." We determined that 13 participants used the median or average forecast, as suggested by statements such as "instead of using the highest degree or the lowest degree, I used an average of the two and used as much of the middle of both as I could figure". Six participants used the area of some part of the p-box to make their decision (e.g., "I'm looking at the shaded area below 32 if it's approximately at least 25% of the shaded part I send blankets") and five participants used the upper bound (e.g., "If the bottom line was below 10% then I wasn't sending the blankets"). In addition, 1 participant in the pbox condition used the range in the probability of freezing, provided by the two bounds.

In the ensembles condition, 23 participants used the median or average, 20 participants used the lower bound, and 1 used the upper bound. In addition, ten participants described using the majority of forecasts to make their decision (e.g. "I just decided if the bulk of the forecasts were 30 percent or lower I would not send blankets") and eight participants used the range make their decision (e.g. "If there was a big gap for example, the lowest was 10% and the highest was 70% id say yes the alpacas need blankets"). The remaining participants (23 in the ensembles condition and 21 in the p-box condition) gave responses that were unclear in terms of determining which aspect of the visualisation they were using to make their decision or explicitly stated that they had no strategy or were guessing.

When we were able to deduce a crossover point (41 responses in ensembles and 42 responses in p-box), we found that most participants did not use the rational crossover point of 0.2. In the ensembles condition, 12 participants used 0.2 as their crossover point, while 28 participants used a value above 0.2 and 1 participant used a value below 0.2. In the p-box condition, only 5 participants used 0.2 as their crossover point, 33 participants used a value above 0.2, and 4 participants used a value below 0.2 (see `supplement ▶ qualitative-analysis ▶ qualitative-analysis.xlsx` for a complete breakdown).

5.2 Experiment 2

As a few participants reported using the majority of forecasts to make their decisions, we decided to look at whether *how forecasts are distributed* could impact participants' decision-making strategies using ensemble representations. The results of Experiment 2 show the estimates for the α (-1.83; 95% CI: [-2.19, -1.51]) and β (2.93, 95% CI: [2.62, 3.26]) parameters are very similar to the estimates for Experiment 1 (see `supplement ▶ R ▶ 04-analysis.Rmd` for details). Further, in the ensembles skewed right condition, the estimated value γ was 0.14 (95% CI: [0.08, 0.22]), suggesting that participants in the ensembles skewed right condition likely almost exclusively use the lower bound of the forecasts to make their decisions (Figure 6). On the other hand, the estimate of γ is both

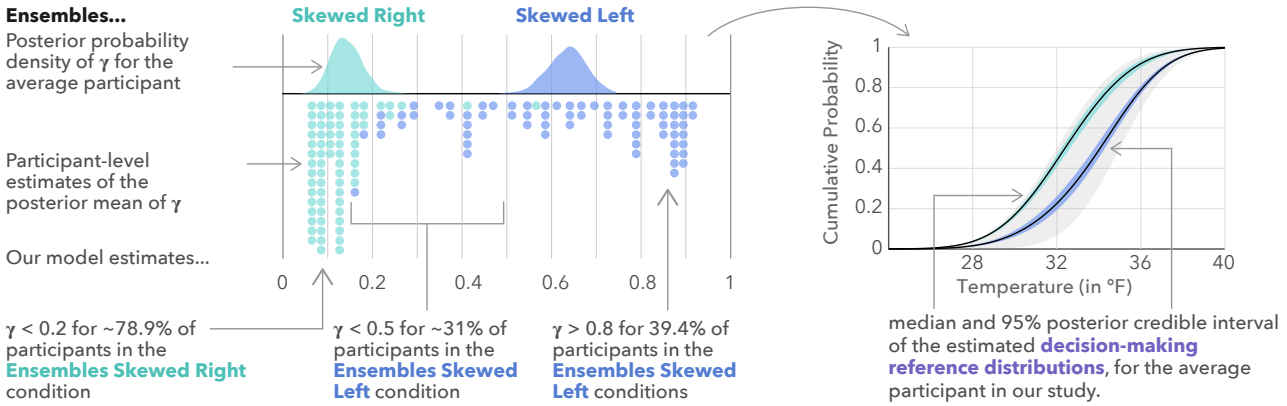


Figure 6: The main result of Experiment 2. We present the posterior probability estimates of the optimism parameter γ for the two conditions (left), and the estimated decision-making reference distribution (right).

much larger in magnitude, and somewhat less precise (0.64; 95% CI: [0.55, 0.71]) for participants in the ensembles skewed left condition (Figure 6).

This slightly greater variance in the estimate of γ in the ensembles skewed left condition is likely attributable to participants adopting a wider range of strategies (which correspond to different values of γ). Following Vuore et al. [63], we use our model to calculate the proportion of participants in each condition who are estimated to be above or below a certain threshold for γ. Our model estimates the mean value of γ to be less than 0.2 for approximately 78.9% (95% CI: [64.8%, 90.1%]) of the participants in the ensembles skewed right condition. Conversely, the model estimates the mean value of γ to be greater than 0.8 for approximately 39.4% (95% CI: [29.6%, 50.7%]) of the participants, and estimates γ to be less than 0.5 for approximately 31% (95% CI: [21.1%, 39.4%] of the participants in the ensembles skewed left condition. To put these values into context, if participants adopted the averaging strategy, after dismissing the best-case forecast—the forecast which predicts the lowest probability of freezing—in the ensembles skewed right condition (or, the worst-case forecast in the ensembles skewed left condition), the values of γ would be 0.17 (or 0.85). This suggests that most participants in the ensembles skewed right condition are most likely not taking the best-case forecast into account, and a large proportion of participants in the ensembles skewed left condition are likely not taking the worst-case forecast into account.

In our qualitative analysis, we found that most participants made their decision based on the part of the visualisation containing the most forecasts, either by using the median/average (18 participants in ensembles skewed right and 17 participants in ensembles skewed left) or by using the majority/consensus forecast (23 participants in ensembles skewed right and 23 participants in ensembles skewed left). In the ensembles skewed right condition, 11 participants reported using the lower bound and 2 participants used the upper bound as their primary strategy. In the ensembles skewed left condition, 8 participants reported using the lower bound and 4 participants reported using the upper bound. Of the responses from which we were able to deduce a crossover point (39 in ensembles skewed right and 34 in ensembles skewed left), only a few participants used the optimal crossover point of 20% (5 in ensembles skewed right and 4 in ensembles skewed left). In contrast, the majority of participants (28 in ensembles skewed right and 24 in ensembles skewed left) used some crossover point greater than 20%, while some participants (6 in ensembles skewed right and 6 in ensembles skewed left) chose a crossover point below 20%.

5.3 Experiment 3

We conducted Experiment 3 as a robustness check to ensure that the accidental difference in phrasing ("reliable" vs "equally reliable") between the p-box and ensembles conditions in Experiment 1 did not impact those results. The difference in the estimate of γ parameter between the two phrasings in Experiment 3 was 0.002 (95% CI: [-0.13, 0.14]) indicating that the phrasing has little-to-no impact on participants responses. In addition, we found the estimate for the β (2.23, 95% CI: [1.93, 2.56]) parameter to be relatively similar to those of the previous two experiments; the estimate of α (-0.90; 95% CI: [-1.24, -0.57]) suggests that the average participant might be somewhat less biased compared to the previous experiments. The results of this analysis can be found in `supplement ▶ R ▶ 04-analysis.Rmd`.

6 Discussion

6.1 How Do People Interpret Visualisations of Multiple Forecasts?

Our work investigated whether *how multiple forecasts are visualised* impacts the strategies that the decision-makers adopt. Prior work [22, 49] has argued that ensembles may be more likely to lead to viewers failing to distinguish between the two types of uncertainty, and treating each forecast to be *equally likely*. Thus, we expected participants to primarily use an averaging approach ($\gamma \approx 0.5$) to decision-making when forecasts were presented using ensembles. Conversely, p-boxes were proposed as a representation for more accurately communicating (whilst also distinguishing between) probabilistic uncertainty and incertitude simultaneously. As p-boxes only present the bounds of incertitude, we expected participants in our experiment to make decisions largely based on

one of these bounds (i.e., $\gamma \approx 0$ or $\gamma \approx 1$); further, assuming that most people are uncertainty-averse, we expected more participants to make decisions based on the lower bound than the upper bound.

Contrary to our expectations, the results of Experiment 1 found that, for *uniformly distributed forecasts*, the type of visualisation seems to have little difference in participants' decision-making strategies (Figure 5C-D). We found that p-boxes and ensembles were equally likely to lead to decisions which place a much greater weight on the lower bound than the upper bound of the forecasts, and our qualitative analysis of participants' self-reported strategies corroborated this finding. However, the results of Experiment 2 suggest that *how the forecasts are distributed* can impact participants' decision-making strategies in the case of ensembles (which would obviously not occur in the case of p-boxes as the visualisation would be unaffected by the distribution of forecasts as long as the lower and upper bounds remain unchanged). Specifically, we found that if a majority of the forecasts either clustered close to the lower or upper bound, participants likely made decisions based on the cluster of forecasts (Figure 6), but placed some weight on the worst-case scenario (which was only relevant in ensembles skewed right).

Our results on ensembles are not too dissimilar from those of Padilla et al. [38]. In their study [38], participants, when shown multiple forecasts as ensembles, predicted a trend which was roughly somewhere between the median and the worst-case forecasts on average. The results from our quantitative analysis are also supported by some of our qualitative responses. In Experiment 1, more participants in the ensembles condition reported making a decision based on the average or majority/consensus forecasts (33 in ensembles vs 13 in p-boxes); similarly, the participants in Experiment 2 also mostly made their decision based on the average or majority/consensus forecasts.

Our results can help designers make more informed choices regarding which representation to use. If a designer wants a decision-maker to make decisions which are *not* influenced by how the forecasts are distributed, p-boxes might be the better design choice. If p-boxes are used, the designer can expect the decision-maker to place greater weight on the worst-case forecast; however, the typical participant will likely incorporate additional factors besides the worst-case forecast into their decision. While our results suggests that the use of ensembles leads to decision-makers adopting similar strategies if the set of forecasts are approximately uniformly distributed, there is the possibility that there may be more variance in the decision-making strategies that viewers adopt when presented with ensembles as it surfaces more information. On the other hand, if a designer wants decisions to be made based on the average or median forecast, and the forecasts *are clustered* around the median or average forecast, they could possibly use ensembles. While none of the approaches evaluated in this study for representing multiple forecasts will likely lead to the typical participant making decisions solely based on the average or median forecasts, we believe that highlighting or other approaches of emphasising the median forecast in an ensemble representation could potentially be more effective in eliciting the desired average-based decision-making strategy. This problem requires more design exploration, which we hope to explore in future work.

6.2 Transparency, Expressiveness and Effectiveness

Ensemble representations are more transparent and expressive [36] compared to p-boxes. In fact, one of the arguments put forth in favor of p-boxes is that by being less expressive, they suppress potentially unnecessary information (the individual forecasts) and draw the viewers' attention to the information which a designer might consider more relevant for decision-making under ambiguity (the bounds). Moreover, it could be argued that p-boxes may be a more effective representation than ensembles in certain scenarios where the lower bound is important and there may be a clustering of forecasts. On the other hand, we do not know if p-boxes would be considered less trustworthy than ensembles by participants due to being less transparent. This highlights a potential trade-off between the expressiveness and effectiveness of these representations.

We can push this notion of reducing expressiveness even further to create visual representations which show only the lower bound (or some other aggregate of the forecasts which the designer considers to be most relevant to the decision maker). Alternatively, a designer might create a visualisation that strikes a balance between ensembles and p-boxes, perhaps by making the individual forecasts less visually salient compared to the bounds. Based on the results of our current study, the consequences of such design choices remain unclear and would require further investigation.

6.3 Potential Applications to Communicating Multiverse Analysis

In multiverse analyses [58], every reasonable analysis for a given dataset and research question is implemented. Multiverse analyses surface the uncertainty in a result due to both the statistical variability in the modeling process as well as due to arbitrary—but justifiable—choices that researchers typically make in a data analysis process (which is referred to as possibilistic uncertainty [22, 49]). Communicating the results of a multiverse analysis [22, 31, 49, 50] represents *one* scenario where a viewer should be able to distinguish between probabilistic uncertainty and possibilistic uncertainty (i.e., incertitude). Prior work [49] has used a variant of the p-box—a representation constructed using the upper and lower bounds of consonance curves [1, 5, 42, 45, 56, 61, 64] instead of CDFs—predicated on the argument that this representation could more likely lead to the desired *possibilistic* interpretation.

A *possibilistic interpretation* of a multiverse result means that any conclusions should be based on which results are possible. For instance, if the goal is to determine whether a treatment effect is positive or not (e.g., a one-tailed t-test), then there are two *possibilistic interpretations* of the result—if the lower bound of the estimates does not overlap zero at a specific confidence level, one can conclude that the treatment is *definitely* positive at that confidence level; if only the upper bound of the estimates does not overlap zero, one can conclude that the treatment's effect is *possibly* positive at that confidence level. These two interpretations map onto the maximin and maximax decision rules outlined in §3. However, drawing conclusions based on the latter interpretation is likely unwise, as it will inflate already existing concerns of potential false positive findings in scientific studies.

Thus, the challenge of any visual representation lies in ensuring that the reader of a multiverse analysis interprets the uncertainty due to choices in the analysis process in the desired *possibilistic* sense. Additionally, we want to avoid letting them fall into the trap of an incorrect *probabilistic interpretation*—considering each individual analysis as equally likely [22, 49] and adopting the *principle of indifference* for drawing conclusions about the overall result. Our results suggest that p-boxes are likely a better-suited, but still imperfect, candidate for communication in this context. P-boxes will likely lead to somewhat conservative decisions and are less sensitive to how the individual distributions are distributed. However, p-boxes, without additional training or instructions, would likely lead to viewers performing some form of weighting between the upper and lower bound of the distributions instead of making decisions based only on the lower bound.

6.4 Expression of Concern for Alpacas Not Reflected in Decisions

Across the two experiments, at least 23 participants reported that they were cautious in their decisions because they did not want alpacas to die, compared to only one participant who reported being cautious because they wanted a bonus. In fact, one participant said "I'd rather keep the alpacas alive than receive a bonus" and another said "I value life and suffering of animals over money", implying that they would act even more cautiously than the rational decision point.[6]

This outpouring of concern for the alpacas—or even broader examples of risk and uncertainty-averse behaviour—were, however, not always reflected in people's actual decisions. For instance, one participant stated that "I always chose the most conservative and safe choice. So if the percentage of likelihood that the temp fell below 32 was about 40% or higher, then I would send it", implying that a crossover point of 0.4 is conservative. However, the crossover point would need to be below 0.2 to be more conservative than is rational. Similarly, another participant reported: "if the chance was significant of 32° or below weather (30% or higher) I would send blankets. I did not want alpacas to die and to lose 5000 dollars." In fact, based on the qualitative responses, the crossover point was greater than the optimal (of 0.2) for 70% of the participants for whom we could deduce a crossover point, which is also reflected in the model estimates (see Figure 5). While we do not have any suggestions on how to address this issue, we found the disconnect between what participants reported they were doing and what they actually did quite interesting. It is possible that participants may be less sensitive to incentives in such tasks than previously anticipated [6], even with both a monetary and emotional incentive; recent work has called for such theories to be tested in HCI and visualisation [51].

6.5 Limitations

Our study methodology has other limitations in addition to potential issues with incentives discussed above. We used an online survey, which allowed us to collect data from a large number of participants and make inferences about the impact of visual representations on average. Participant responses in surveys may involve guessing, and self-reported data may introduce additional noise. This may have driven some of the inconsistencies between participants' multiple choice responses (where they indicated which forecast they primarily used to make decisions) and their textual explanations of the strategy they used to perform the task.

In our study, participants only performed a binary decision-making task—whether the temperature was going to be below freezing based on one or multiple forecasts. However, in reality, decision-making can be more complex and may involve choosing between three or more alternatives—e.g., the decision to bet on a soccer game would involve three possible outcomes (team A wins, team B wins or the game ends in a draw)—or may not even be possible to be distilled into discrete choices. We hope to explore the impact of representations on more complex decision-making tasks in future work.

[6]Our favorite response was from a pilot participant, who said "I don't want to kill any alpacas because I'm not a psychopath. [...] If the overall average (visually) was over 40ish percent I gave those stinky creatures blankets. To be honest, bonuses are usually a sham but the thought of killing creatures made me want to do good things in virtual survey alpaca world."

7 Conclusion

In this paper, we wanted to understand how people make decisions when faced with multiple forecasts distributions regarding an event. We conducted two experiments where we compared: (1) p-boxes and uniformly distributed ensembles; and (2) ensembles where a majority of the forecasts are clustered close to the upper or lower bound. In addition, we conducted a third experiment as a robustness check. We found that, for the average participant, multiple forecasts represented as p-boxes and ensembles will likely lead to similar decisions where greater weight is placed on the worst-case forecast. However, in the case of ensemble representations, we also found that participants' decisions will be likely sensitive to how the individual forecasts are distributed. Our findings suggests that designers of multiple forecast visualisations who wish their users to adopt particular normative strategies for decision-making under epistemic uncertainty must carefully consider the strategies their visualisation designs may induce in their users.

Acknowledgments

This research is supported by NSF 2211939. We would like to thank Alex Kale, Fumeng Yang, and members of MUCollective at Northwestern University for their thoughtful feedback on this research. We also thank the anonymous reviewers for their comments and feedback.

References

[1] Valentin Amrhein and Sander Greenland. 2022. Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. *Journal of Information Technology* 37, 3 (Sept. 2022), 316–320. https://doi.org/10.1177/02683962221105904 Publisher: SAGE Publications Ltd.

[2] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological methods* 10 (Dec. 2005), 389–96. https://doi.org/10.1037/1082-989X.10.4.389

[3] Ryan Best and Jay Boice. [n.d.]. Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree | projects.fivethirtyeight.com. https://projects.fivethirtyeight.com/covid-forecasts/. [Accessed 02-09-2024].

[4] Michael Betancourt. 2020. *Towards A Principled Bayesian Workflow.* https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

[5] Allan Birnbaum. 1961. A Unified Theory of Estimation, I. *The Annals of Mathematical Statistics* 32, 1 (1961), 112–135. https://www.jstor.org/stable/2237612 Publisher: Institute of Mathematical Statistics.

[6] Colin F. Camerer and Robin M. Hogarth. 1999. *The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework.* Springer Netherlands, Dordrecht, 7–48. https://doi.org/10.1007/978-94-017-1406-8_2

[7] Graciela Chichilnisky. 2000. An axiomatic approach to choice under uncertainty with catastrophic risks. *Resource and Energy Economics* 22, 3 (2000), 221–231. https://doi.org/10.1016/S0928-7655(00)00032-4

[8] Samantha R Cook, Andrew Gelman, and Donald B Rubin. 2006. Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics* 15, 3 (2006), 675–692. https://doi.org/10.1198/106186006X136976

[9] Jonathan Cox, Donald House, and Michael Lindell. 2013. Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification* 3, 2 (2013).

[10] Daniel Ellsberg. 1961. Risk, Ambiguity, and the Savage Axioms*. *The Quarterly Journal of Economics* 75, 4 (11 1961), 643–669. https://doi.org/10.2307/1884324 arXiv:https://academic.oup.com/qje/article-pdf/75/4/643/5399198/75-4-643.pdf

[11] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173718

[12] Scott Ferson and Jack Siegrist. 2012. Verified Computation with Probabilities. In *Uncertainty Quantification in Scientific Computing*, Andrew M. Dienstfrey and Ronald F. Boisvert (Eds.). IFIP Advances in Information and Communication Technology, Vol. 377. Springer Berlin Heidelberg, Berlin, Heidelberg, 95–122. https://doi.org/10.1007/978-3-642-32677-6_7

[13] National Center for Advancing Translational Sciences. [n.d.]. OpenData Portal | SARS-CoV-2 Variants & Therapeutics Therapeutic Activity Explorer. https://opendata.ncats.nih.gov/covid19/variant/activity. [Accessed 02-09-2024].

[14] Camilla Bretteville Froyn. 2005. Decision Criteria, Scientific Uncertainty, and the Globalwarming Controversy. *Mitigation and Adaptation Strategies for Global Change* 10, 2 (April 2005), 183–211. https://doi.org/10.1007/s11027-005-3782-9

[15] Jonah Gabry, Rok Češnovar, Andrew Johnson, and Steve Bronder. 2024. *cmdstanr: R Interface to 'CmdStan'*. https://mc-stan.org/cmdstanr/ R package version 0.8.0, https://discourse.mc-stan.org.

[16] Andrew Gelman. [n.d.]. Florida. Comparing Economist and Fivethirtyeight forecasts. https://statmodeling.stat.columbia.edu/2020/08/27/florida-comparing-economist-and-fivethirtyeight-forecasts/. [Accessed 02-09-2024].

[17] Andrew Gelman. 2024. Polling averages and political forecasts and what do you really think is gonna happen in November? https://statmodeling.stat.columbia.edu/2024/07/18/polling-averages-and-political-forecasts-and-what-do-you-really-think-is-gonna-happen-in-november. [Accessed 02-09-2024].

[18] Gerd Gigerenzer, Ralph Hertwig, Eva Van Den Broek, Barbara Fasolo, and Konstantinos V. Katsikopoulos. 2005. "A 30% Chance of Rain Tomorrow": How Does the Public Understand Probabilistic Weather Forecasts? *Risk Analysis* 25, 3 (2005), 623–629. https://doi.org/10.1111/j.1539-6924.2005.00608.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2005.00608.x

[19] Itzhak Gilboa and David Schmeidler. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 2 (Jan. 1989), 141–153. https://doi.org/10.1016/0304-4068(89)90018-9

[20] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174079

[21] Peter Gärdenfors and Nils-Eric Sahlin. 1982. Unreliable Probabilities, Risk Taking, and Decision Making. *Synthese (Dordrecht)* 53, 3 (1982), 361–386.

[22] Brian D. Hall, Yang Liu, Yvonne Jansen, Pierre Dragicevic, Fanny Chevalier, and Matthew Kay. 2022. A Survey of Tasks and Visualizations in Multiverse Analysis Reports. *Computer Graphics Forum* 41, 1 (2022), 402–426. https://doi.org/10.1111/cgf.14443

[23] Leonid Hurwicz. 1951. The generalized Bayes minimax principle: a criterion for decision making under uncertainty. *Cowles Comm. Discuss. Paper Stat* 335 (1951), 1950.

[24] Harald Ibrekk and M. Granger Morgan. 1987. Graphical Communication of Uncertain Quantities to Nontechnical People. *Risk Analysis* 7, 4 (1987), 519–529. https://doi.org/10.1111/j.1539-6924.1987.tb00488.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.1987.tb00488.x

[25] Alex Kale, Matthew Kay, and Jessica Hullman. 2021. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 272–282. https://doi.org/10.1109/TVCG.2020.3030335

[26] Josh Katz. [n.d.]. 2016 Election Forecast: Who Will Be President? (Published 2016) | nytimes.com. https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html#other-forecasts. [Accessed 02-09-2024].

[27] Peter Klibanoff, Massimo Marinacci, and Sujoy Mukerji. 2005. A Smooth Model of Decision Making under Ambiguity. *Econometrica* 73, 6 (2005), 1849–1892. https://doi.org/10.1111/j.1468-0262.2005.00640.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2005.00640.x

[28] Frank Knight. 1921. *Risk, Uncertainty and Profit.* Hart, Schaffner & Marx, Boston.

[29] Andreas Lange. 2003. Climate Change and the Irreversibility Effect – Combining Expected Utility and MaxiMin. *Environmental and Resource Economics* 25, 4 (Aug. 2003), 417–434. https://doi.org/10.1023/A:1025054716419

[30] Robert J. Lempert and Myles T. Collins. 2007. Managing the Risk of Uncertain Threshold Responses: Comparison of Robust, Optimum, and Precautionary Approaches. *Risk Analysis* 27, 4 (2007), 1009–1026. https://doi.org/10.1111/j.1539-6924.2007.00940.x

[31] Le Liu, Lace Padilla, Sarah H. Creem-Regehr, and Donald H. House. 2019. Visualizing Uncertain Tropical Cyclone Predictions using Representative Samples from Ensembles of Forecast Tracks. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 882–891. https://doi.org/10.1109/TVCG.2018.2865193 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[32] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics* 27, 2, 1753–1763. https://doi.org/10.1109/TVCG.2020.3028985

[33] Massimo Marinacci. 2002. Probabilistic Sophistication and Multiple Priors. *Econometrica* 70, 2 (2002), 755–764. http://www.jstor.org/stable/2692290

[34] David McInerney, Robert Lempert, and Klaus Keller. 2012. What are robust strategies in the face of uncertain climate threshold responses? *Climatic Change* 112, 3 (June 2012), 547–568. https://doi.org/10.1007/s10584-011-0377-1

[35] Martin Modrák, Angie H. Moon, Shinyoung Kim, Paul Bürkner, Niko Huurre, Kateřina Faltejsková, Andrew Gelman, and Aki Vehtari. 2023. Simulation-Based Calibration Checking for Bayesian Computation: The Choice of Test Quantities Shapes Sensitivity. *Bayesian Analysis* 1, 1 (Jan. 2023). https://doi.org/10.1214/23-ba1404

[36] Tamara Munzner. 2014. *Visualization Analysis and Design.* A K Peters/CRC Press, Boca Raton, Florida, Chapter 5, 94 – 114.

[37] Limor Nadav-Greenberg and Susan L. Joslyn. 2009. Uncertainty Forecasts Improve Decision Making Among Nonexperts. *Journal of Cognitive Engineering and Decision Making* 3, 3 (2009), 209–227. https://doi.org/10.1518/155534309X474460 arXiv:https://doi.org/10.1518/155534309X474460

[38] Lace Padilla, Racquel Fygenson, Spencer C. Castro, and Enrico Bertini. 2023. 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 (2023), 12–22. https://doi.org/10.1109/TVCG.2022.3209457

[39] 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 (Oct. 2017), 40. https://doi.org/10.1186/s41235-017-0076-1

[40] Lace M. K. Padilla, Maia Powell, Matthew Kay, and Jessica Hullman. 2021. Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations. *Frontiers in Psychology* 11 (2021). https://doi.org/10.3389/fpsyg.2020.579267

[41] Martin Peterson. 2017. *An Introduction to Decision Theory* (2 ed.). Cambridge University Press.

[42] C Poole. 1987. Beyond the confidence interval. *American Journal of Public Health* 77, 2 (Feb. 1987), 195–199. https://doi.org/10.2105/AJPH.77.2.195

[43] Kristin Potter, Paul Rosen, and Chris R. Johnson. 2012. *From Quantification to Visualization: A Taxonomy of Uncertainty Visualization Approaches.* IFIP Advances in Information and Communication Technology, Vol. 377. Springer Berlin Heidelberg, Berlin, Heidelberg, 226–249. https://doi.org/10.1007/978-3-642-32677-6_15

[44] R Core Team. 2024. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[45] Zad Rafi and Sander Greenland. 2020. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology* 20, 1 (Sept. 2020), 244. https://doi.org/10.1186/s12874-020-01105-9

[46] Frank P. Ramsey. 2016. *Truth and Probability.* Springer International Publishing, Cham, 21–45. https://doi.org/10.1007/978-3-319-20451-2_3

[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 (April 2016), 154–172. https://doi.org/10.1080/13875868.2015.1137577

[48] Jibonananda Sanyal, Song Zhang, Jamie Dyer, Andrew Mercer, Philip Amburn, and Robert Moorhead. 2010. Noodles: A Tool for Visualization of Numerical Weather Model Ensemble Uncertainty. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1421–1430. https://doi.org/10.1109/TVCG.2010.

181

[49] Abhraneel Sarma, Kyle Hwang, Jessica Hullman, and Matthew Kay. 2024. Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 607, 15 pages. https://doi.org/10.1145/3613904.3642375

[50] Abhraneel Sarma, Alex Kale, Michael Jongho Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. 2023. multiverse: Multiplexing Alternative Data Analyses in R Notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 148, 15 pages. https://doi.org/10.1145/3544548.3580726

[51] Abhraneel Sarma, Sheng Long, Michael Correll, and Matthew Kay. 2022. Tasks and Telephones: Threats to Experimental Validity due to Misunderstandings of Visualisation Tasks and Strategies. In *2024 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*.

[52] Abhraneel Sarma, Xiaoying Pu, Yuan Cui, Michael Correll, Eli T Brown, and Matthew Kay. 2024. Odds and Insights: Decision Quality in Exploratory Data Analysis Under Uncertainty. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1034, 14 pages. https://doi.org/10.1145/3613904.3641995

[53] L. J. Savage. 1951. The Theory of Statistical Decision. *J. Amer. Statist. Assoc.* 46, 253 (1951), 55–67. https://doi.org/10.1080/01621459.1951.10500768

[54] Sonia Savelli and Susan Joslyn. 2013. The Advantages of Predictive Interval Forecasts for Non-Expert Users and the Impact of Visualizations. *Applied Cognitive Psychology* 27, 4 (2013), 527–541. https://doi.org/10.1002/acp.2932 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2932

[55] Nicholas Shackel. 2007. Bertrand's Paradox and the Principle of Indifference. *Philosophy of Science* 74, 2 (2007), 150–175. https://doi.org/10.1086/519028

[56] Kesar Singh, Minge Xie, and William E. Strawderman. 2007. Confidence Distribution (CD): Distribution Estimator of a Parameter. *Lecture Notes-Monograph Series* 54 (2007), 132–150. https://www.jstor.org/stable/20461464 Publisher: Institute of Mathematical Statistics.

[57] David Spiegelhalter. 2017. Risk and Uncertainty Communication. *Annual Review of Statistics and Its Application* 4, Volume 4, 2017 (2017), 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148

[58] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* 11, 5 (2016), 702–712.

[59] S. S Stevens. 1957. On the psychophysical law. *Psychological review* 64, 3 (1957), 153–181.

[60] S. S. (Stanley Smith) Stevens and Geraldine. Stevens. 1975. *Psychophysics : introduction to its perceptual, neural, and social prospects*. Transaction Books, New Brunswick, U.S.A.

[61] Kevin M. Sullivan and David A. Foster. 1990. Use of the Confidence Interval Function. *Epidemiology* 1, 1 (1990), 39–42. https://www.jstor.org/stable/20065621 Publisher: Lippincott Williams & Wilkins.

[62] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. 2020. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv:1804.06788 [stat.ME] https://arxiv.org/abs/1804.06788

[63] Matti Vuorre, Matthew Kay, and Niall Bolger. 2024. Communicating causal effect heterogeneity. https://doi.org/10.31234/osf.io/mwg4f

[64] Min-ge Xie and Kesar Singh. 2013. Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statistical Review* 81, 1 (2013), 3–39. https://doi.org/10.1111/insr.12000 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12000.

[65] Fumeng Yang, Maryam Hedayati, and Matthew Kay. 2023. Subjective Probability Correction for Uncertainty Representations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 836, 17 pages. https://doi.org/10.1145/3544548.3580998

[66] Fumeng Yang, Chloe Rose Mortenson, Erik Nisbet, Nicholas Diakopoulos, and Matthew Kay. 2024. In Dice We Trust: Uncertainty Displays for Maintaining Trust in Election Forecasts Over Time. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 389, 24 pages. https://doi.org/10.1145/3613904.3642371

[67] Hang Zhang and Laurence T Maloney. 2012. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience* 6 (2012), 21111.