Leffrang, Dirk; Müller, Oliver

Visualizing Uncertainty in Time Series Forecasts: The Impact of Uncertainty Visualization on Users' Confidence, Algorithmic Advice Utilization, and Forecasting Performance

Journal of Forecasting

Journal of Forecasting

WILEY

RESEARCH ARTICLE OPEN ACCESS

Visualizing Uncertainty in Time Series Forecasts: The Impact of Uncertainty Visualization on Users' Confidence, Algorithmic Advice Utilization, and Forecasting Performance

Dirk Leffrang | Oliver Müller

Department of Information Systems, Paderborn University, Paderborn, Germany

Correspondence: Dirk Leffrang (dirk.leffrang@uni-paderborn.de)

Received: 10 May 2023 | **Revised:** 27 May 2024 | **Accepted:** 23 October 2024

Keywords: advice utilization | decision-making | human-centered computing | uncertainty visualization

ABSTRACT

Time series forecasts are always associated with uncertainty. However, experimental studies on the impact of uncertainty communication provide inconclusive results on the effect of providing this uncertainty to end users. In this study, we examine the impact of uncertainty visualizations on advice utilization in the context of time series forecasts with line charts. Based on a literature review, we identified probabilistic framing versus frequency framing as a theoretical foundation for studying the topic. We then used the Judge Advisor System (JAS) as a framework to create an experimental design with two treatments (95% prediction interval [PI] and ensemble plots), one control group (point plot), and various mediating variables (e.g., confidence, graph literacy, and domain knowledge). The results of an online experiment ($N = 239$) indicate a U-shaped relation between uncertainty visualization and forecasting performance. Additionally, we examine how confidence, advice utilization, and other factors mediate the effect of uncertainty visualizations. This paper highlights the benefits of PI plots for researchers and practitioners engaged in the development of effective uncertainty visualizations for decision-making processes.

1 | Introduction

Whether tracking the development of a pandemic or predicting stock market developments, time series forecasts are ubiquitous. Such forecasts are inherently uncertain, particularly when the prediction horizon is long. Incorporating uncertainty improves forecasting performance in multiple scenarios (e.g., Cepni, Guney, and Swanson 2020; Liang et al. 2021). Thus, from a normative perspective, uncertainty should be made transparent to users (van der Bles et al. 2020) and incorporated into decision-making processes (Tannert, Elvers, and Jandrig 2007).

However, humans frequently employ simplifying heuristics, rather than a rational process that weighs all available information (Kahneman 2011). One illustrative example is the salience bias, whereby individuals tend to give greater weight to prominent visualization properties, such as colors or dynamics, relative to other properties (Taylor 1982). This can result in decision-makers failing to consider crucial uncertainty information, frequently represented by less salient properties such as shading.

The existing literature reports mixed results regarding the impact of uncertainty visualizations on human decision-making. On the one hand, incorporating information about uncertainty into probability distribution forecasts, such as weather forecasts, leads to more transparency (e.g., Padilla, Kay, and Hullman 2022). On the other hand, individuals often misinterpret uncertainty information. For instance, they may erroneously assume that uncertainty information is deterministic rather than probabilistic (Joslyn and Savelli 2021).

One potential explanation for these inconclusive findings is that researchers have employed a multitude of experimental designs to study the phenomenon. For instance, some studies have focused on trust or confidence as dependent variables (e.g., Padilla et al. 2022), while others have used advice utilization (e.g., McGrath et al. 2023) or forecasting performance (e.g., Hoekstra et al. 2014). Furthermore, the results of existing experiments are inconsistent, with some studies reporting a positive effect (e.g., Fernandes et al. 2018), while others found that uncertainty visualizations can confuse participants (e.g., Zhou et al. 2017). A more systematic approach is necessary to analyze the problem domain effectively. Therefore, we aim to explore the following research questions:

1. Do uncertainty visualizations affect the forecasting performance of users for algorithmic time series forecasts?
2. Does this effect depend on users' confidence and algorithmic advice utilization?

We conducted an online experiment using the Judge Advisor System (JAS) to investigate our research questions. The JAS is a standard approach for studying advice-taking in psychological behavioral research (Bonaccio and Dalal 2006). It has been effective in other research domains, such as algorithm aversion (e.g., Prahl and Swol 2017; Logg et al. 2019). The JAS provides Weight of Advice (WOA) as a dependent variable and a standardized experimental procedure (see Bonaccio and Dalal 2006).

In our experiment, participants were asked to forecast the number of hospitalizations due to COVID-19 over a period of three weeks by manually forecasting the trajectory of a historical time series. We presented 239 participants with several time series of the number of occupied hospital beds due to COVID-19. Subsequently, we provided them with advice in the form of an algorithmic forecast. The control group received advice without uncertainty information, while the two experimental groups received advice with different types of uncertainty visualizations, namely, a 95% prediction interval (PI) plot and an ensemble plot. Participants were then given the option to adjust their initial prediction. We computed a structural equation modeling (SEM) to estimate the correlative relations of major variables studied in the existing literature on uncertainty visualization and algorithmic advice-taking.

PI plots (medium perceived uncertainty) were the best choice compared to both point plots (low perceived uncertainty) and ensemble plots (high perceived uncertainty). Thus, our results imply an inverted U-shaped relation between uncertainty visualizations and confidence in the algorithm. In addition, our findings indicate that both confidence and advice utilization mediate the impact of uncertainty visualizations on forecasting performance.

Our research is among the few papers exploring uncertainty visualization in time series forecasting. In addition, this study introduces the JAS framework and SEM to uncertainty visualizations in time series analysis using line charts. Instead of examining variables separately, we conducted a mediation analysis to reveal how variables such as confidence and advice utilization mediate the effect of uncertainty visualizations on forecasting performance. Based on our results, practitioners should use PI uncertainty visualizations, if their goal is to maximize confidence in algorithms, advice utilization from them, and performance in collaboration with them in forecasting under uncertainty.

2 | Related Work

This section presents a summary of our literature review on different uncertainty visualization formats. More specifically, this paper studies three types of uncertainty visualization: probabilistic framing, frequency framing, and visualizations without explicit uncertainty information. Figure 1 provides a conceptual overview of how these uncertainty visualizations impact confidence, advice utilization, and forecasting performance. Most prior studies use these constructs as dependent variables. Table 1 presents an overview of the related work, which we will discuss in depth in the following subsections.

2.1 | Probabilistic Framing

Probabilistic uncertainty visualizations, such as PI plots, are among the most commonly utilized uncertainty representations. PI plots have gained increasing popularity in both media and research. For instance, 60% of COVID-19 visualizations with line charts included a PI (Zhang, Sun, and Padilla 2021).

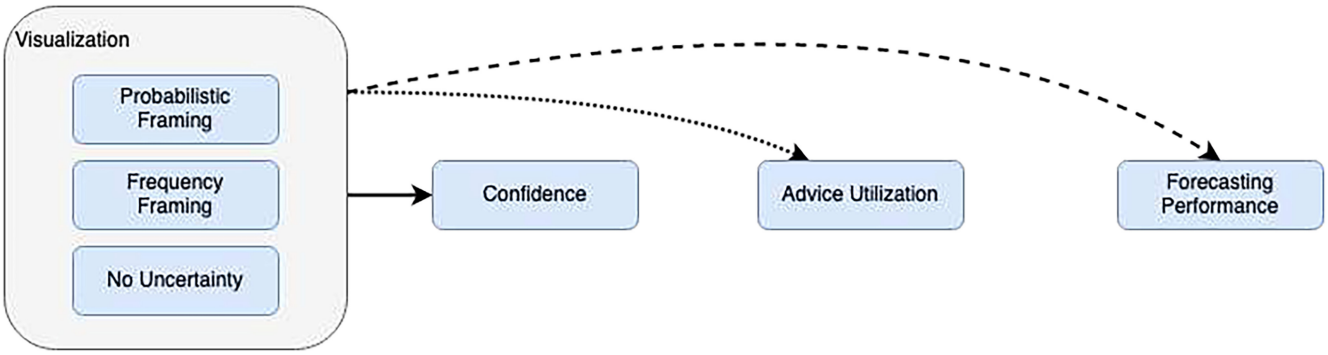

FIGURE 1 | Factors influencing the forecasting process.

TABLE 1 | Related work.

Study	Context	Participants	Probabilistic format	Frequency format	No uncertainty	DV
Belia et al. (2005)	None	473	✓	✓		FP
Fernandes et al. (2018)	Logistics	408	✓	✓	✓	AU
Greis et al. (2016)	Weather	38	✓		✓	AU
Grounds, Joslyn, and Otsuka (2017)	Weather	644; 388	✓		✓	AU
Hoekstra et al. (2014)	None	560	✓			FP
Ibrekk and Morgan (1987)	Weather	45	✓			FP
Joslyn, Nemec, and Savelli (2013)	Weather	302; 312	✓		✓	C; AU
Kale, Kay, and Hullman (2021)	Sports	622	✓	✓	✓	AU
Karduni et al. (2021)	Health	92; 212; 267	✓	✓	✓	AU
Kim et al. (2019)	Health	50; 800; 300	✓	✓	✓	AU
Leffrang and Müller (2021)	Health	84	✓	✓	✓	AU
Liu et al. (2017)	Weather	201		✓		C; FP
McGrath et al. (2023)	Rent	95	✓		✓	AU
Padilla, Ruginski, and Creem-Regehr (2017)	Weather	182; 43	✓	✓		AU
Padilla et al. (2022)	Health	200; 299; 800	✓	✓	✓	C; FP
Padilla et al. (2022)	Health	1200; 675; 675	✓	✓	✓	AU
Ruginski et al. (2016)	Weather	200	✓	✓	✓	FP
Tak, Toet, and Van Erp (2014)	Geology	140; 24	✓	✓		AU
Toet et al. (2019)	Health	64	✓	✓		FP
Zhou et al. (2017)	Maintenance	42	✓		✓	C

Abbreviations: AU, advice utilization; C, confidence; DV, dependent variable; FP, forecasting performance.

Several studies have demonstrated that the provision of probabilistic uncertainty information has a positive effect on human judgment and decision-making. For example, when forecasting time-constant rental prices, participants were more likely to utilize advice when they observed the uncertainty distribution of the advice (McGrath et al. 2023). In the context of COVID-19, PI plots increased advice utilization compared to other visualizations when forecasting hospitalization rates (Leffrang and Müller 2021). Displaying probabilistic uncertainty information increased the willingness to use advice for a spatial weather-related task compared to no uncertainty being visualized (Joslyn, Nemec, and Savelli 2013). Finally, in a web-based agricultural game, participants demonstrated a subjective preference for PI plots over other uncertainty representations (Greis et al. 2016).

Nevertheless, while probabilistic uncertainty visualizations, such as PI plots, can be beneficial, there is also evidence that people may misinterpret this information. A common mistake is the tendency of individuals to perceive values within an uncertainty interval as accurate and values outside of it as erroneous (Correll and Gleicher 2014). They incorrectly perceive interval boundaries as partitions of the probabilistic space into categorically different areas of true and false values (Padilla et al. 2018), concluding that closed areas contain points of the same probability (Ibrekk and Morgan 1987). Surprisingly, this issue is not exclusive to those lacking statistical expertise. Even those with statistical training may err in their interpretation of predictive intervals (Hoekstra et al. 2014; Belia et al. 2005).

2.2 | Frequency Framing

Frequency-based visualizations like ensemble plots represent a distinct form of visual uncertainty. They are generated by varying the input parameters of a forecasting model and combining different predicted future trajectories in a single visualization (Liu et al. 2017; Stephenson and Doblas-Reyes 2000). This approach differs from probabilistic interval visualizations, which are simpler by design. However, although ensemble plots divert the attention through the benefit of more information, the visual system can offset this drawback by aggregating noisy local features and calculating summary statistics (Alvarez and Oliva 2008).

The theoretical foundation of ensemble plots is rooted in the frequency format hypothesis (Gigerenzer 1996) that posits that individuals process and comprehend information more effectively when presented in frequency rather than numerical or probabilistic formats. This is because humans are more likely to encounter frequencies than probabilities in their daily lives (Gigerenzer 1996). Frequency-based visualizations like ensemble plots can mitigate misconceptions about uncertainty intervals, such as misinterpreting the PI of spatial hurricane path forecasts as the magnitude of the storm (Ruginski et al. 2016). In another study using spatial data, ensemble and interval plots were both effective for most participants if the objective was to achieve the greatest perceived certainty towards the center of a range and decreasing probabilities towards the outer edges. However, ensemble plots demonstrated the best fit for this goal (Tak, Toet, and Van Erp 2014).

Empirical evidence also indicates that frequency format visualizations lead to more advice utilization than other visualization types. For instance, ensemble plots of incidence values increased the perception of risk associated with COVID-19 compared to probabilistic visualizations like PI plots and plots without an explicit visualization of uncertainty (Padilla et al. 2022). In another study, nonspecialist participants viewed various visualizations of children's growth distributions and estimated the probability that a child would reach a certain height. Ensemble plots achieved the smallest biases compared to probabilistic visualizations (Toet et al. 2019).

Empirical studies also found that individuals tend to associate ensemble plots with more perceived uncertainty. When evaluating COVID-19 visualizations, participants rated their confidence in PI plots higher than in ensemble plots. However, these participants performed poorer when forecasting the anticipated shift (e.g., rising, declining, remaining constant, or uncertain) in the rate of COVID-19 fatalities over the upcoming fortnight (Padilla et al. 2022). Interestingly, including an excessive number of lines in ensemble plots did not enhance participants' forecasting abilities but undermined their trust in the visualization (Padilla et al. 2022). In a sports study, participants demonstrated comparable performance with probabilistic and frequency visualizations under low variance. However, with high variance, probabilistic displays were more effective in detecting differences in performance distributions (Kale, Kay, and Hullman 2021). While ensembles align perceptions of hurricane risk with actual risk, ensemble plots can sometimes appear cluttered and disorganized, colloquially known as "spaghetti plots", which can hinder accurately interpreting hurricane paths (Liu et al. 2017).

2.3 | Summary

Despite growing research on the impact of uncertainty visualizations on human decision-making, it remains unclear whether they have a positive impact overall. In general, individuals typically do not incorporate all information given in a task and often incorporate statistical information incorrectly (Kim et al. 2019). For instance, researchers found that uncertainty visualizations increased confidence for low cognitive load tasks but decreased confidence for high cognitive load tasks in the context of predictive maintenance (Zhou et al. 2017). In another study on predicting geospatial hurricane courses, participants misinterpreted probabilistic boundaries of PIs and overweighted individual lines of ensemble plots (Padilla, Ruginski, and Creem-Regehr 2017). Consequently, to account for the possibility that uncertainty visualizations have no effect or even negative effects, researchers have argued for the inclusion of a "no uncertainty visualization" condition (e.g., a point plot) in experimental studies on uncertainty visualizations (see Hullman et al. 2019).

In conclusion, researchers have observed both positive and negative effects of uncertainty visualizations with a focus on spatial data, time-constant data, or specific dependent variables. The objective of this study is to investigate the effect of uncertainty visualizations on different variables of the forecasting process in the context of time series forecasting with line graphs. To this end, we employ a holistic design, integrating various dependent variables studied in prior experiments and the pathways connecting them. Due to the inconclusive results of previous studies, we chose not to preregister specific hypotheses. Instead, we opted for a more holistic and exploratory investigation of the relations between different variables.

3 | Experimental Design

3.1 | Research Model

The objective of our study was to investigate whether the exposure to different uncertainty visualizations affects forecasting performance for algorithmic time series forecasts and whether this effect depends on the level of confidence and advice utilization. To this end, we employed the JAS paradigm to assess the extent to which individuals rely on advice from diverse sources (Sniezek and Buckley 1995). In a typical JAS, a person acts as a judge and makes a decision. Initially, the judge forms her own opinion. Next, the judge can consult the opinion of another person or algorithm representing the advisor. Finally, the judge weighs the extent to which she follows the advisor's recommendation (Bonaccio and Dalal 2006).

To study the effect of uncertainty visualizations on multiple dependent variables, we use SEM. A SEM is a statistical model representing the relations between multiple variables in a complex system as a set of equations. SEM allows for the simultaneous analysis of multiple relations among variables, which can provide a more comprehensive understanding of a system than traditional regression analysis. Graphically, arrows connect the variables in the model to represent these relations. SEM can be used to test hypotheses about causal relations, estimate the strength and direction of these relations, and evaluate the overall fit of the model to the data (Hayes 2022).

Our research model is depicted in Figure 2. The treatment consisted of three different visualizations of uncertainty, namely, a 95% PI as a representative of probabilistic framing, an ensemble plot as a representative of frequency framing, and a line chart only showing point plots without any uncertainty representation (control condition serving as the baseline).

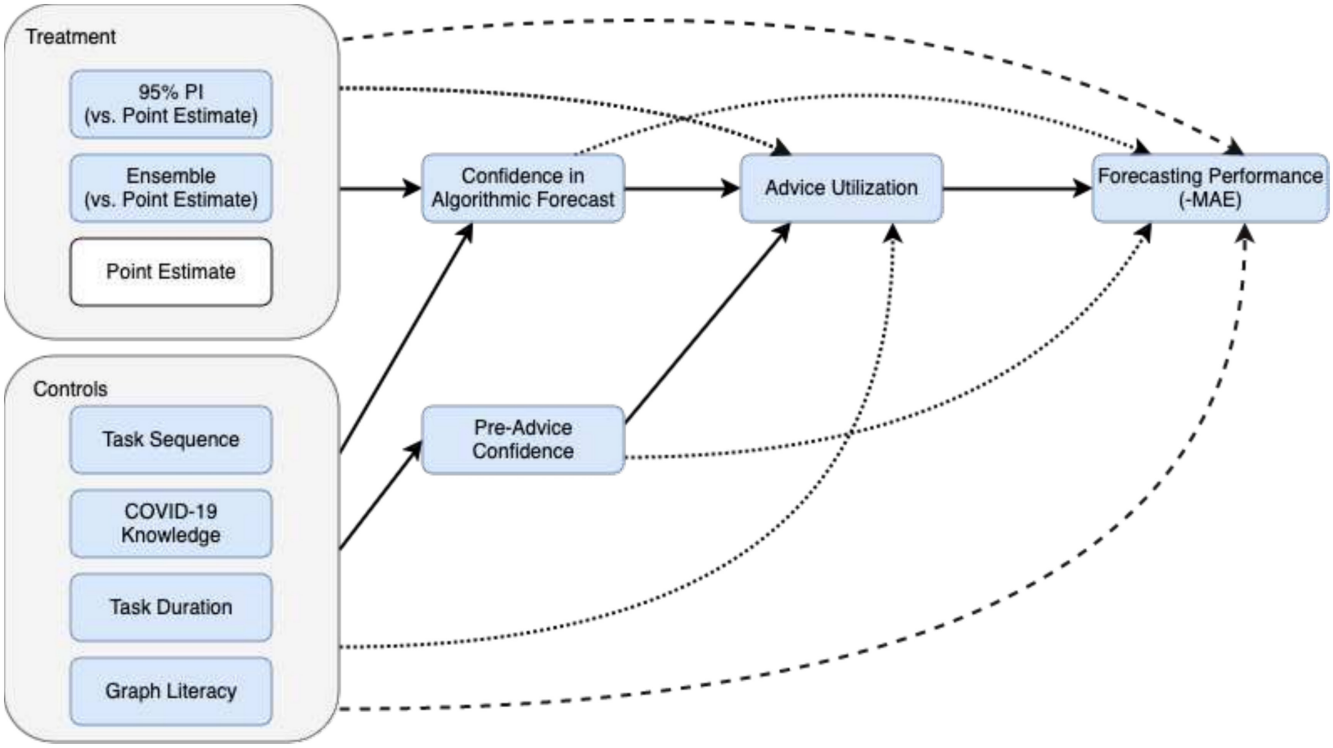

FIGURE 2 | Research model. Different lines were chosen for visual simplicity.

Confidence is a typical mediator variable in the JAS and can refer to the confidence of the advisor or the judge (Bonaccio and Dalal 2006). The operational definition of confidence is an anticipation of the degree to which a judgment is correct (Bonaccio and Dalal 2006). Prior studies found that people's self-confidence influences the acceptance of algorithmic suggestions more than confidence in the algorithmic advisor (Chong et al. 2022). Better information on the part of the judge or an aversion towards algorithms may result in noncompliance with algorithmic advice (Sun et al. 2021). Consequently, we measured the confidence in both the algorithmic forecast and the judge's pre-advice confidence.

We used the WOA variable to measure advice utilization, which takes the value of 0 if the initial estimate remains unadjusted. The greater the discrepancy between the final estimate and the initial estimate, the higher the value of WOA. Consistent with previous research, we applied Winsorization to all WOA values that exceeded 1 (e.g., Logg et al. 2019).

$$WOA = \frac{|final\ forecast - initial\ forecast|}{|advisor's\ forecast - initial\ forecast|} \quad (1)$$

In addition, we statistically adjusted for several factors that could influence the results, including domain knowledge, graph literacy, confidence, task duration, and task sequence. Prior experiments suggest that domain knowledge influences advice utilization from algorithms (Logg et al. 2019). Therefore, we included the 13 COVID-19 knowledge questions of Azlan et al. (2020). We coded all correct answers as 1 and all false answers as 0. The sum of all answers indicated the domain expertise of the participant. Graph literacy is the ability to understand information presented graphically (Galesic and Garcia-Retamero 2011). We used the short graph literacy (SGL) scale to determine the graph literacy of participants (Okan et al. 2019). Task duration was centered and measured in seconds. It controlled for effort in processing the tasks. We centered it and included it to control for effort in processing the tasks. Task sequence controlled for fatigue.

Additionally, the negated mean absolute error (MAE) between the actual values and the participants' final forecasts measured the forecasting performance. Given the three forecasts for each task, we averaged the variables over the three forecasts. It is important to note that the axis scaling can influence the size of the estimated treatment effect (Hofman, Goldstein, and Hullman 2020). Therefore, we kept the ordinate constant. We standardized all variables by centering them to have a mean of 0 and scaling them to have a standard deviation of 1. This procedure allows for comparing coefficients between variables on different scales (Hayes 2022).

3.2 | Task and Procedures

We used the Coronavirus (COVID-19) hospitalization rate for different countries as the application context. The exact time period and the country names were anonymized. The experiment consisted of two steps:

1. Participants viewed a graph that showed the daily COVID-19 hospitalization rates for each country over a period of 16 weeks. They made forecasts for the next 3 weeks, specifically for $t + 7$ (i.e., in 7 days), $t + 14$, and $t + 21$. Figure 3 shows an annotated version of the first step, which we provided to ensure a basic understanding of time series charts.

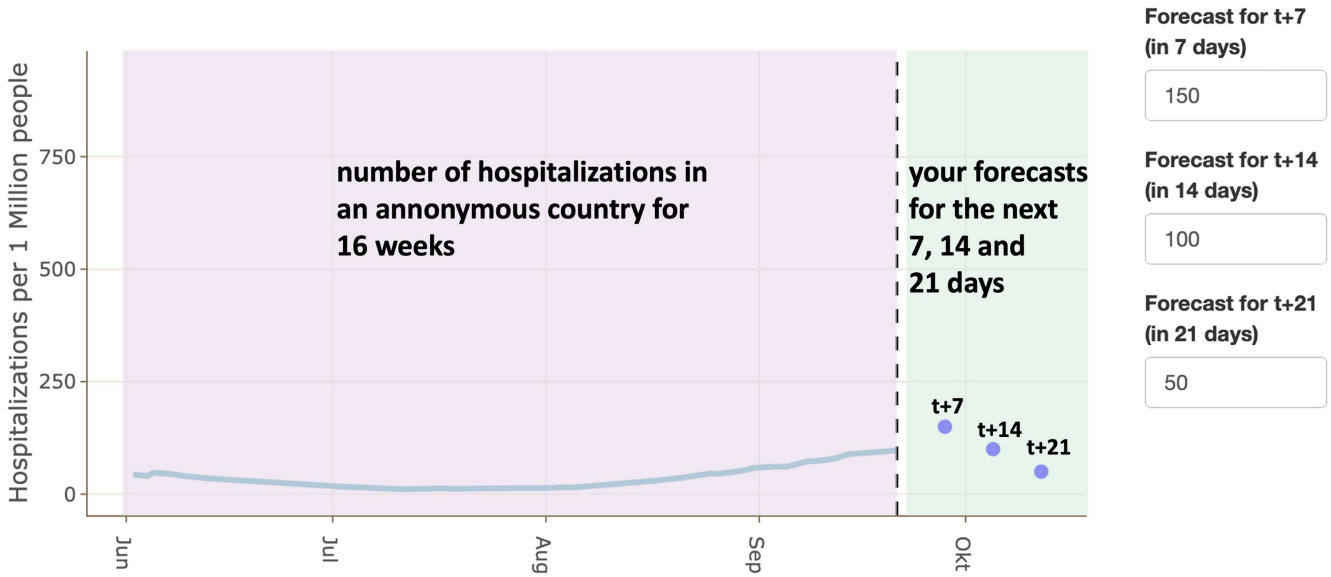

FIGURE 3 | Annotated example of a user interface in the first step.

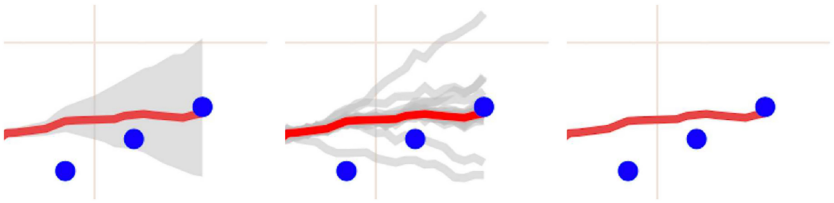

FIGURE 4 | 95% PI, ensemble, and point plot for the forecast of COVID-19 hospitalizations.

The visualization interactively shifted the points representing the predictions.

2. We then provided participants with an algorithmic prediction. Participants now had the opportunity to adjust their initial forecast.

Each participant repeated these steps for nine countries. We randomly shuffled the order of the countries and excluded each participant's first round (warm-up trial) from the analysis. This enabled participants to familiarize themselves with the user interface and tasks.

3.3 | Conditions and Participants

Figure 4 illustrates examples of the three conditions. We used Facebook's Prophet R package, which provides fully automated Bayesian additive models that can represent nonlinear trends and seasonal effects, to generate the forecasts (Taylor and Letham 2018). We selected a 95% PI plot for the probabilistic framing. Prophet employs Markov Chain Monte Carlo (MCMC) simulations to generate samples from the posterior distribution of the model parameters, from which the PI is derived based on quantiles (Facebook 2024). We created ensembles by randomly selecting 20 forecasts from the MCMC samples. The control group received a point plot representing no visual uncertainty. As a central line can influence decision-making (Kale, Kay, and Hullman 2021), we included a central line in all conditions.

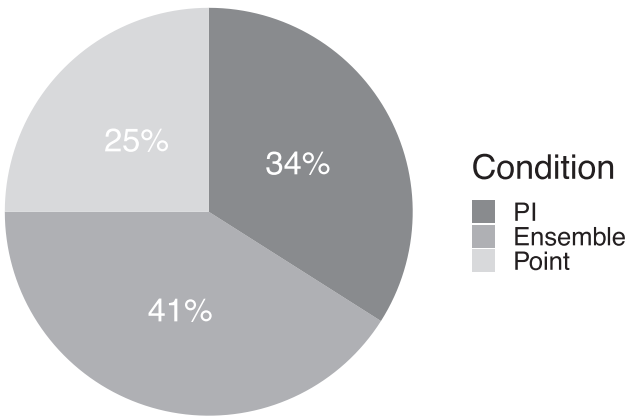

FIGURE 5 | Most perceived uncertainty for 95% PI, ensemble, and point plot.

We conducted a manipulation check with an additional sample of 100 participants from Prolific to assess whether the perceived uncertainty actually differed between our conditions. Participants had to choose the plot showing the most perceived uncertainty. As illustrated in Figure 5, most participants associated ensemble plots with the highest perceived uncertainty, followed by PI and point plots. Furthermore, we used an 11-point

Likert scale to measure the perceived uncertainty. The ensemble plot received the highest mean perceived uncertainty score of 6.070 (sd: 2.508), while the mean score for the 95% PI was 5.800 (sd: 2.314). The point condition received the lowest mean perceived uncertainty score of 5.610 (sd: 1.989).

Between February 24 and July 6, 2022, 272 Prolific participants from the UK and the United States, balanced on gender, participated in the between-subjects experiment. They received £2.50 as compensation. We excluded participants with incomplete observations and those who were too fast (less than 90 s) or slow (more than 2 standard deviations above the average). Furthermore, we excluded observations for which the dependent variable was not defined, that is, if participants had the same initial forecast as the algorithm. A total of 239 participants remained in the final data set.

4 | Results

4.1 | Model-Free Evidence

We collected a total of 1877 observations representing task-participant combinations. Table 2 summarizes participants' key demographic and dependent variables (i.e., WOA and MAE).

Figure 6 displays the distributions of the WOA and MAE by condition. The PI plot condition (medium perceived uncertainty) has a marginally higher median WOA value (0.188) compared to the ensemble plot (high perceived uncertainty; 0.168) and a higher median WOA value compared to the point plot (low perceived uncertainty; 0.081). Regarding the median values for the MAE, the PI condition is associated with a marginally lower error (29.333) in comparison to the ensemble plot (32.000) and the point plot (32.667). Therefore, the model-free results indicate that the PI condition is associated with marginally higher advice utilization and marginally higher forecasting performance.

On average, the MAE between the algorithmic predictions and the ground truth values (23.700) is 53% lower than the MAE between the participants' initial predictions and the ground truth values (50.800). This indicates that the algorithmic forecasts are superior to the initial human forecasts and suggests that adhering to the model's advice should increase the overall forecasting performance. The MAE of the participants' final predictions (36.920) is 27% lower on average than the MAE of their initial predictions (50.800). This suggests that the participants benefited from utilizing the algorithmic advice.

TABLE 2 | Overall summary statistics.

		Mean		
N	**WOA**	**MAE**	**Age**	**Female (%)**
239	0.278	39.279	37.610	47.774

4.2 | Estimation Results

Table 3 provides a comprehensive summary of the results. We employed clustered standard errors to account for correlated errors due to the repeated measures within individuals. Figure 7 illustrates the results of the SEM in an aggregated way. Each black arrow represents a significant coefficient from the mediation analysis. Gray arrows represent insignificant paths.

4.2.1 | Confidence in Algorithmic Forecast

The PI condition is associated with increased confidence in the algorithmic forecast by 0.192 standard deviations compared to the no uncertainty condition (point plot). In contrast, the results reveal no significant effect for the second condition, that is, the ensemble plot ($p \geq 0.1$). None of the control variables is significantly associated with confidence in the algorithmic forecast ($p \geq 0.1$).

4.2.2 | Pre-Advice Confidence

Regarding pre-advice confidence, the results indicate a significant association of 0.129 standard deviations more confidence per standard deviation of task duration ($p < 0.05$). For the other control variables, the results show no significant relations.

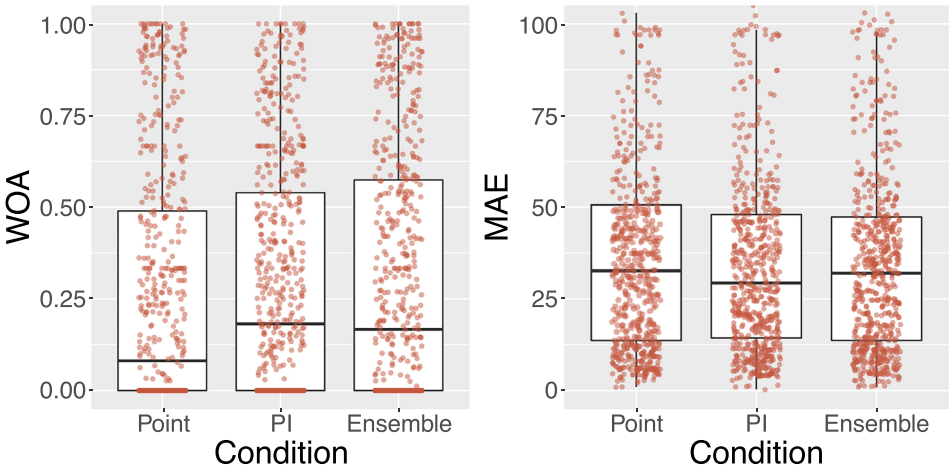

FIGURE 6 | Boxplots for WOA and MAE by conditions. We limited the ordinate axis in the MAE plot to 100.

TABLE 3 | Structural equation modeling (SEM) results.

	Estimate	Std. err.	*z*-value	$P(>\|z\|)$
Forecasting performance (-MAE) ˜				
PI	−0.010	0.078	0.125	0.900
Ensemble	0.049	0.062	−0.787	0.431
Task sequence	−0.005	0.009	0.559	0.576
COVID-19 knowledge	0.025	0.019	−1.325	0.185
Task duration	0.029	0.028	−1.043	0.297
Graph literacy	0.163	0.069	−2.403	0.016
WOA	0.163	0.020	−8.172	0.000
Confidence in algorithmic forecast	0.261	0.025	−10.534	0.000
Pre-advice confidence	-0.041	0.023	1.801	0.072
Advice utilization (WOA) ˜				
PI	0.117	0.095	1.236	0.216
Ensemble	0.088	0.101	0.871	0.384
Task sequence	−0.005	0.009	−0.608	0.543
COVID-19 knowledge	−0.001	0.023	−0.042	0.966
Task duration	0.177	0.041	4.312	0.000
Graph literacy	−0.058	0.061	−0.945	0.344
Confidence in algorithmic forecast	0.129	0.035	3.675	0.000
Pre-advice confidence	−0.206	0.034	−6.111	0.000
Confidence in algorithmic forecast ˜				
PI	0.192	0.112	1.718	0.086
Ensemble	−0.035	0.114	−0.311	0.756
Task sequence	−0.002	0.010	−0.245	0.807
COVID-19 knowledge	0.031	0.026	1.184	0.236
Task duration	0.082	0.052	1.594	0.111
Graph literacy	0.044	0.070	0.618	0.536
Pre-advice confidence ˜				
Task sequence	−0.000	0.008	−0.007	0.995
COVID-19 knowledge	0.028	0.027	1.043	0.297
Task duration	0.129	0.050	2.561	0.010
Graph literacy	−0.083	0.087	−0.951	0.342
Intercepts:				
MAE	0.431	0.231	1.866	0.062
WOA	0.040	0.253	0.157	0.875
Confidence in algorithmic forecast	−0.395	0.282	−1.402	0.161
Pre-advice confidence	−0.179	0.306	−0.585	0.558
Variances:				
MAE	0.881	0.070	12.606	0.000
WOA	0.927	0.041	22.880	0.000
Confidence in algorithmic forecast	0.981	0.053	18.457	0.000
Pre-advice confidence	0.980	0.066	14.848	0.000

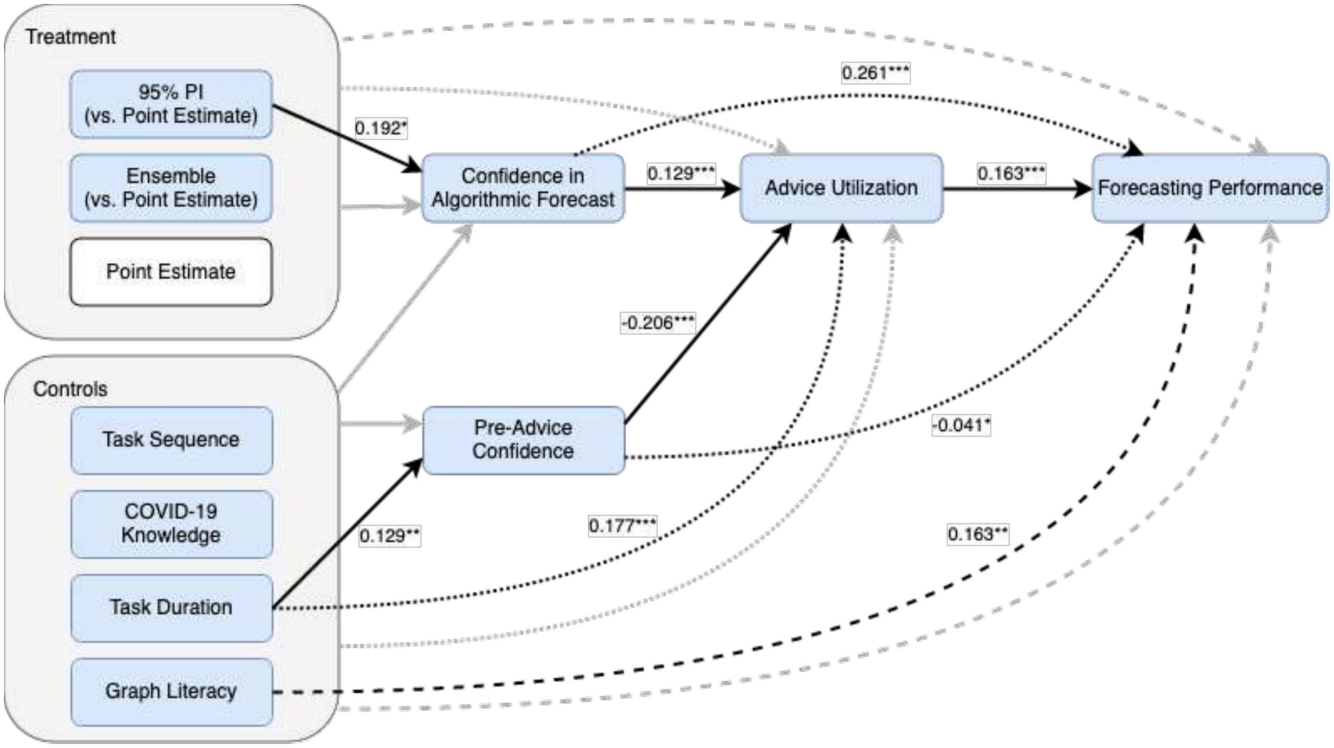

FIGURE 7 | Structural equation modeling (SEM) results. Nonsignificant relations have been grayed out for visual simplicity. Coefficients are scaled to changes in the standard deviations. Significance levels: * 90%, ** 95%, *** 99%.

4.2.3 | Advice Utilization (WOA)

Compared to the point plot, the analysis reveals no significant difference in the *direct* association of the conditions PI plot or ensemble plot with advice utilization ($p \geq 0.1$). However, our SEM indicates that an increase in confidence in the algorithmic forecast by one standard deviation is associated with an increase of 0.129 standard deviations in advice utilization ($p < 0.01$). Furthermore, the results suggest a significant decline in advice utilization for each standard deviation increase in pre-advice confidence (-0.206, $p < 0.01$). Regarding our controls, the coefficient for the task duration indicates a significant increase in advice utilization (0.177, $p < 0.01$). Task sequence, COVID-19 knowledge, and graph literacy do not yield significant coefficients in the results.

4.2.4 | Forecasting Performance (-MAE)

Once more, our estimation results do not suggest a significant *direct* association between the treatments and forecasting performance ($p \geq 0.1$). However, according to our model higher confidence in the algorithmic forecast is associated with higher forecasting performance (0.261, $p < 0.01$). In contrast, our results suggest a negative correlation between pre-advice confidence and forecasting performance ($p < 0.01$). Furthermore, our findings indicate that a higher weight of algorithmic advice is associated with higher forecasting performance (0.163, $p < 0.01$). This provides support for the beneficial association between taking advice from algorithms and forecasting performance. Regarding our controls, the SEM reveals a significant association between graph literacy and the final forecasting performance (0.163, $p < 0.05$). The results do not indicate significant coefficients for the task sequence, COVID-19 knowledge, and task duration ($p \geq 0.1$).

Our analysis of the model suggests a path of significant relations from the 95% PI plot treatment to the MAE. The results indicate a positive association between 95% PI plots and confidence in algorithmic advice, which in turn is associated with higher levels of advice utilization and, subsequently, higher forecasting performance. Interestingly, the findings indicate that pre-advice confidence levels are associated with decreased forecasting performance, both directly and indirectly through advice utilization.

4.3 | Post Hoc Analysis

As described earlier, people often misinterpret visual boundaries in plots and perceive these boundaries, which are actually probabilistic, in a categorical manner (Padilla et al. 2018). This may lead to the erroneous interpretation that points lying within a PI or between the lines of an ensemble are "correct" and points outside are "incorrect" (Padilla, Kay, and Hullman 2022). As a consequence, people might adopt a strategy in which they adjust their initial forecasts in such a way that their estimates lie inside the PI or between the most extreme lines of an ensemble plot. This could, in turn, reduce the frequency of large forecast errors.

For point plots, 19.0% of the final forecasts are more than two standard deviations away from the true value. In contrast, the frequency of such outliers is lower for the ensemble (15.4%) and PI condition (14.5%). A Pearson's chi-squared test of independence

indicates significant differences ($p < 0.1$). These results suggest that people may indeed move their initial estimates inside the PI or between the lines of an ensemble plot, reducing the frequency of large forecast errors.

Finally, we examined the robustness of our results by replacing the negated MAE with the negated root mean square error (RMSE). The RMSE is naturally more sensible to large forecast errors. However, all relations remain on similar significance levels, except for the relation between pre-advice confidence and forecasting performance, which is not significant for RMSE as the dependent variable ($p > 0.1$).

5 | Discussion and Conclusion

This paper examines the impact of uncertainty visualizations on forecasting performance in time series forecasts. It highlights the importance of considering uncertainty in decision-making and how confidence, advice utilization, and other factors impact the decision-making process. The paper makes two contributions. First, it consolidates research on uncertainty visualizations by using the JAS framework from psychological behavioral research for the special case of time series forecasts. Second, we go beyond a simple examination of the effect of uncertainty visualizations by using SEM to examine the involvement of other variables, such as confidence and advice utilization.

PI plots representing *probabilistic framing* played a special role among the selected visualization conditions, as presented in Figure 7 in the previous chapter. This visualization form led to a greater degree of confidence in the algorithmic forecast. Given the finding that advice utilization was generally beneficial for final decision-making, this suggests that PI plots were the best choice in the application context. This could help to explain why most COVID-19 visualizations used PI plots (Zhang, Sun, and Padilla 2021). Additionally, we can confirm the beneficial impact of PI plots compared to no uncertainty visualizations (e.g., Joslyn, Nemec, and Savelli 2013; Leffrang and Müller 2021).

Ensemble plots as instances of *frequency framing* may be easier to understand than probabilistic framing (Gigerenzer 1996). However, increasing the perceived level of uncertainty may not always be the optimal approach. Our findings indicate an inverted U-shaped relation between perceived uncertainty and advice utilization. Individuals appear to be more willing to take advice from an algorithm when the algorithm makes them aware of some uncertainty associated with the forecast (in our case: 95% PI plots). However, when confronted with overly effective uncertainty visualizations (in our case: ensemble plots), individuals may become alienated and less likely to take the advice. This points to a possible uncertainty paradox, where increased uncertainty awareness leads to reduced confidence in useful advice, which leads to suboptimal decision-making outcomes.

We cannot confirm the beneficial impacts of ensembles, like mitigating misconceptions about uncertainty (Ruginski et al. 2016), improved risk perception (Padilla et al. 2022), or reduction of biases (Toet et al. 2019) to translate to confidence, advice utilization, or forecasting performance. Thus, we can confirm that visualization formats minimizing human bias are not necessarily optimal for decision-making (Kale, Kay, and Hullman 2021). One explanation could be that the number of forecasts (20) undermined participants' trust (Padilla et al. 2022). Additionally, disorganized ensembles could have confused participants (Liu et al. 2017).

However, both PI and ensemble plots reduced the frequency of very large forecastíng errors, which appears to contradict confusion as the primary explanation. One explanation can be the misconception of visual boundaries as categorical ones (Padilla et al. 2018). This misconception can lead to a misunderstanding of statistical properties (e.g., Hoekstra et al. 2014; Belia et al. 2005). Contrary to prior research, we present a beneficial implication of such a misconception.

Confidence in algorithmic forecast mediated the relation between visualization type and forecasting performance, which we did not observe as a direct one. Thus, we extend prior research on forecasting performance (e.g., Belia et al. 2005; Padilla et al. 2022) with a more nuanced view of the decision-making process, which may explain mixed findings on uncertainty visualization evaluations. Confidence in the algorithmic forecast had a direct positive relation with forecasting performance, according to our results. Additionally, it had a positive indirect relation with a positive relation between advice utilization and forecasting performance. Since the algorithmic advice was generally beneficial, as indicated by the model-free evidence, confidence in its ability seems reasonable to increase forecasting performance. Based on these results, we can confirm the importance of confidence in the algorithmic advisor (Bonaccio and Dalal 2006; Sun et al. 2021).

For **advice utilization**, measured by *WOA*, besides confidence in the algorithmic forecast, duration had a positive direct relation with WOA. However, duration also had a negative relation with WOA mediated by pre-advice confidence according to our results. Thus, our results indicate that more time can increase advice utilization and self-confidence. Self-confidence itself appears to decrease advice utilization according to our results. This is in line with research on overconfidence, which is generally harmful to an effective decision-making process (Bonaccio and Dalal 2006).

For **forecasting performance**, measured by the *negated MAE*, besides confidence in the algorithmic forecast and advice utilization, pre-advice confidence and graph literacy had a direct positive relation to it. For pre-advice confidence, similar to advice utilization, overconfidence may explain these results (Moore and Healy 2008). As graph literacy describes the ability to understand the information in a graph (Galesic and Garcia-Retamero 2011), more graph literacy increasing the forecasting performance appears reasonable.

Overall, this paper provides valuable insights into the impact of uncertainty visualizations on advice utilization in time series forecasts. Data scientists must decide (i) whether to include uncertainty information in their results and, if so, (ii) how to represent it. Our findings indicate that while information about uncertainty can increase confidence in forecasts, it can also be misunderstood, negatively impacting decision-making

and creating an uncertainty paradox. The JAS principle and mediation analyses provide a comprehensive approach to investigating the effect of uncertainty visualizations on advice utilization. PI plots can be a useful tool in achieving an uncertainty balance, as they provide a range of possible outcomes while maintaining clarity. This paper is useful for researchers and practitioners interested in developing effective uncertainty visualizations helping individuals to make more informed decisions and avoid the pitfalls of the uncertainty paradox.

6 | Limitations and Outlook

Our research is not without limitations. First, although WOA is the most frequently used dependent variable in algorithmic advice utilization research, it is subject to certain constraints, including undefined values, the absence of negative values, and the potential for overshooting (Bonaccio and Dalal 2006).

Second, we randomly drew lines from the posterior distribution to differentiate ensemble plots from the more structured PI plots. Future work could examine scenario-reduction techniques to combine the desired properties of each visualization form (e.g., Liu et al. 2019).

Third, we focused on static visualization techniques as they are widely applicable across all media and are most commonly used in the considered use case, based on the media examples we identified. Future research may study animated or interactive visualizations, which might have the capability to convey more information. Furthermore, future work can examine the impact of insights from temporal correlations in forecasts of trajectories that probabilistic or no uncertainty visualizations do not capture.

Finally, the anchoring of initial values on the value in t may have impacted the predictions for $t + 7$, $t + 14$, and $t + 21$ before the algorithmic forecast was shown. Future studies could investigate whether anchoring and adjustment may play a role in forecasting tasks.

Although our research design was firmly grounded in theoretical work on the perception of probabilities and frequencies, we believe that it is necessary to develop a more comprehensive understanding of the causal mechanisms linking various types and properties of uncertainty visualizations to different effects. Future work should study how other factors, such as the way graphs are colored or labeled, or the width of uncertainty ranges, influence the perceived predictive uncertainty and try to build an overarching theoretical framework (e.g., Padilla et al. 2022).

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Alvarez, G. A., and A. Oliva. 2008. "The Representation of Simple Ensemble Visual Features Outside the Focus of Attention." *Psychological Science* 19, no. 4: 392–398.

Azlan, A. A., M. R. Hamzah, T. J. Sern, S. H. Ayub, and E. Mohamad. 2020. "Public Knowledge, Attitudes and Practices Towards COVID-19: A Cross-Sectional Study in Malaysia." *PLoS ONE* 15, no. 5: 1–15.

Belia, S., F. Fidler, J. Williams, and G. Cumming. 2005. "Researchers Misunderstand Confidence Intervals and Standard Error Bars." *Psychological Methods* 10, no. 4: 389–396.

Bonaccio, S., and R. S. Dalal. 2006. "Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences." *Organizational Behavior and Human Decision Processes* 101, no. 2: 127–151.

Cepni, O., I. E. Guney, and N. R. Swanson. 2020. "Forecasting and Nowcasting Emerging Market GDP Growth Rates: The Role of Latent Global Economic Policy Uncertainty and Macroeconomic Data Surprise Factors." *Journal of Forecasting* 39, no. 1: 18–36.

Chong, L., G. Zhang, K. Goucher-lambert, K. Kotovsky, and J. Cagan. 2022. "Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice." *Computers in Human Behavior* 127, no. 107018: 1–10.

Correll, M., and M. Gleicher. 2014. "Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error." *IEEE Transactions on Visualization and Computer Graphics* 20, no. 12: 2142–2151.

Facebook. 2024. "Prophet." Accessed April 19, 2024. http://facebook.github.io/prophet/.

Fernandes, M., L. Walls, S. Munson, J. Hullman, and M. Kay. 2018. "Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making." In *Proceedings of the Conference on Human Factors in Computing Systems*, 1–12. New York, NY, USA: ACM.

Galesic, M., and R. Garcia-Retamero. 2011. "Graph Literacy: A Cross-Cultural Comparison." *Medical Decision Making* 31, no. 3: 444–457.

Gigerenzer, G. 1996. "The Psychology of Good Judgment: Frequency Formats and Simple Algorithms." *Medical Decision Making* 16, no. 3: 273–280.

Greis, M., P. E. Agroudy, H. Schuff, T. Machulla, and A. Schmidt. 2016. "Decision-Making Under Uncertainty: How the Amount of Presented Uncertainty Influences User Behavior." In *Proceedings of the Nordic Conference on Human-Computer Interaction*, 1–4. New York, NY, USA: ACM.

Grounds, M. A., S. Joslyn, and K. Otsuka. 2017. "Probabilistic Interval Forecasts: An Individual Differences Approach to Understanding Forecast Communication." *Advances in Meteorology* 2017: 3932565.

Hayes, A. F. 2022. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York, NY, USA: Taylor & Francis.

Hoekstra, R., R. D. Morey, J. N. Rouder, and E. J. Wagenmakers. 2014. "Robust Misinterpretation of Confidence Intervals." *Psychonomic Bulletin and Review* 21, no. 5: 1157–1164.

Hofman, J. M., D. G. Goldstein, and J. Hullman. 2020. "How visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results." In *Proceedings of the Conference on Human Factors in Computing Systems*, 1–12. Honolulu, HI, USA: ACM.

Hullman, J., X. Qiao, M. Correll, A. Kale, and M. Kay. 2019. "In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation." *IEEE Transactions on Visualization and Computer Graphics* 25, no. 1: 903–913.

Ibrekk, H., and M. G. Morgan. 1987. "Graphical Communication of Uncertain Quantities to Nontechnical People." *Risk Analysis* 7, no. 4: 519–529.

Joslyn, S., L. Nemec, and S. Savelli. 2013. "The Benefits and Challenges of Predictive Interval Forecasts and Verification Graphics for End Users." *Weather, Climate, and Society* 5, no. 2: 133–147.

Joslyn, S., and S. Savelli. 2021. "Visualizing Uncertainty for Non-Expert End Users: The Challenge of the Deterministic Construal Error." *Frontiers in Computer Science* 2, no. 590232: 1–12.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York, NY, USA: Farrar, Straus and Giroux.

Kale, A., M. Kay, and J. Hullman. 2021. "Visual Reasoning Strategies for Effect Size Judgments and Decisions." *IEEE Transactions on Visualization and Computer Graphics* 27, no. 2: 272–282.

Karduni, A., D. Markant, R. Wesslen, and W. Dou. 2021. "A Bayesian Cognition Approach for Belief Updating of Correlation Judgement Through Uncertainty Visualizations." *IEEE Transactions on Visualization and Computer Graphics* 27, no. 2: 978–988.

Kim, Y. S., L. A. Walls, P. Krafft, and J. Hullman. 2019. "A Bayesian Cognition Approach to Improve Data Visualization." In *Proceedings of the Conference on Human Factors in Computing Systems*, 1–14. New York, NY, USA: ACM.

Leffrang, D., and O. Müller. 2021. "Should I Follow This Model? The Effect of Uncertainty Visualization on the Acceptance of Time Series Forecasts." In *IEEE Workshop on Trust and Expertise in Visual Analytics (TREX)*, 20–26. New Orleans, LA, USA: IEEE.

Liang, C., F. Ma, L. Wang, and Q. Zeng. 2021. "The Information Content of Uncertainty Indices for Natural Gas Futures Volatility Forecasting." *Journal of Forecasting* 40, no. 7: 1310–1324.

Liu, L., A. P. Boone, I. T. Ruginski, et al. 2017. "Uncertainty Visualization by Representative Sampling From Prediction Ensembles." *IEEE Transactions on Visualization and Computer Graphics* 23, no. 9: 2165–2178.

Liu, L., L. Padilla, S. H. Creem-Regehr, and D. H. House. 2019. "Visualizing Uncertain Tropical Cyclone Predictions Using Representative Samples From Ensembles of Forecast Tracks." *IEEE Transactions on Visualization and Computer Graphics* 25, no. 1: 882–891.

Logg, J. M., J. A. Minson, D. A. Moore, and U. States. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151: 90–103.

McGrath, S., P. Mehta, A. Zytek, I. Lage, and H. Lakkaraju. 2023. "When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making." *Transactions on Machine Learning Research*: 1–21.

Moore, D. A., and P. J. Healy. 2008. "The Trouble With Overconfidence." *Psychological Review* 115, no. 2: 502–517.

Okan, Y., E. Janssen, M. Galesic, and E. A. Waters. 2019. "Using the Short Graph Literacy Scale to Predict Precursors of Health Behavior Change." *Medical Decision Making* 39, no. 3: 183–195.

Padilla, L., R. Fygenson, S. C. Castro, and E. 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, no. 1: 12–22.

Padilla, L., H. Hosseinpour, R. Fygenson, J. Howell, R. Chunara, and E. Bertini. 2022. "Impact of COVID-19 Forecast Visualizations on Pandemic Risk Perceptions." *Scientific Reports* 12, no. 1: 1–14.

Padilla, L., M. Kay, and J. Hullman. 2022. "Uncertainty Visualization." In *Computational Statistics in Data Science*, 405–421. Wiley.

Padilla, L. M., S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. 2018. "Decision Making With Visualizations: A Cognitive Framework Across Disciplines." *Cognitive Research: Principles and Implications* 3, no. 1: 1–25.

Padilla, L. M., I. T. Ruginski, and S. H. Creem-Regehr. 2017. "Effects of Ensemble and Summary Displays on Interpretations of Geospatial Uncertainty Data." *Cognitive Research: Principles and Implications* 2, no. 1: 1–16.

Prahl, A., and L. V. Swol. 2017. "Understanding Algorithm Aversion: When Is Advice From Automation Discounted?." *Journal of Forecasting* 36, no. 6: 691–702.

Ruginski, I. T., A. P. Boone, L. M. Padilla, et al. 2016. "Non-Expert Interpretations of Hurricane Forecast Uncertainty Visualizations." *Spatial Cognition and Computation* 16, no. 2: 154–172.

Sniezek, J. A., and T. Buckley. 1995. "Cueing and Cognitive Conflict in Judge-Advisor Decision Making." *Organizational Behavior and Human Decision Processes* 62, no. 2: 159–174.

Stephenson, D. B., and F. J. Doblas-Reyes. 2000. "Statistical Methods for Interpreting Monte Carlo Ensemble Forecasts." *Tellus, Series A: Dynamic Meteorology and Oceanography* 52, no. 3: 300–322.

Sun, J., D. Zhang, H. Hu, and J. A. Van Mieghem. 2021. "Predicting Human Discretion to Adjust Algorithmic Prescription: A Large-Scale Field Experiment in Warehouse Operations." *Management Science* 68, no. 2: 846–865.

Tak, S., A. Toet, and J. Van Erp. 2014. "The Perception of Visual Uncertainty Representation by Non-Experts." *IEEE Transactions on Visualization and Computer Graphics* 20, no. 6: 935–943.

Tannert, C., H. D. Elvers, and B. Jandrig. 2007. "The Ethics of Uncertainty: In the Light of Possible Dangers, Research Becomes a Moral Duty." *EMBO Reports* 8, no. 10: 892–896.

Taylor, S. E. 1982. "The Availability Bias in Social Perception and Interaction." In *Judgment Under Uncertainty: Heuristics and Biases*, 138–140. Cambridge University Press.

Taylor, S. J., and B. Letham. 2018. "Forecasting at Scale." *American Statistician* 72, no. 1: 37–45.

Toet, A., J. B. F. van Erp, E. M. Boertjes, and S. van Buuren. 2019. "Graphical Uncertainty Representations for Ensemble Predictions." *Information Visualization* 18, no. 4: 373–383.

van der Bles, A. M., S. van der Linden, A. L. J. Freeman, and D. J. Spiegelhalter. 2020. "The Effects of Communicating Uncertainty on Public Trust in Facts and Numbers." *Proceedings of the National Academy of Sciences of the United States of America* 117, no. 14: 7672–7683.

Zhang, Y., Y. Sun, and L. Padilla. 2021. "Mapping the Landscape of COVID-19 Crisis Visualizations." In *Proceedings of the Conference on Human Factors in Computing Systems*, pp. 1–23. Yokohama, Japan: ACM.

Zhou, J., S. Z. Arshad, S. Luo, and F. Chen. 2017. "Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making." In *Proceedings of the Conference on Human-Computer Interaction*, 23–39. Bombay, India: Springer.