Should I Follow this Model? The Effect of Uncertainty Visualization on the Acceptance of Time Series Forecasts

Dirk Leffrang* Oliver Müller†

Department of Information Systems, Paderborn University, Germany

ABSTRACT

Time series forecasts are ubiquitous, ranging from daily weather forecasts to projections of pandemics such as COVID-19. Communicating the uncertainty associated with such forecasts is important, because it may affect users' trust in a forecasting model and, in turn, the decisions made based on the model. Although there exists a growing body of research on visualizing uncertainty in general, the important case of visualizing prediction uncertainty in time series forecasting is under-researched. Against this background, we investigated how different visualizations of predictive uncertainty affect the extent to which people follow predictions of a time series forecasting model. More specifically, we conducted an online experiment on forecasting occupied hospital beds due to the COVID-19 pandemic, measuring the influence of uncertainty visualization of algorithmic predictions on participants' own predictions. In contrast to prior studies, our empirical results suggest that more salient visualizations of uncertainty lead to decreased willingness to follow algorithmic forecasts.

Index Terms: uncertainty visualization—visualization—visualization techniques— human-centered computing—decision-making—non-expert audiences

1 INTRODUCTION

Virtually every day we are confronted with time series forecasts. Examples include weather forecasts, predictions of customer demand, the development of financial markets, and the projection of global pandemics like COVID-19. Such forecasts of the future are always associated with uncertainty, especially when prediction horizons are long. From a normative perspective, decision makers should incorporate this uncertainty into their decision-making processes [30]. However, research on behavioral aspects of judgment and decision-making has shown that people are subject to cognitive biases and often apply simplifying heuristics instead of following rational decision theory [13].

Several studies examined the effects of uncertainty visualizations on decision making, but to the best of our knowledge, the case of uncertainty in time series forecasts has not yet been investigated. Focusing on spatial data, Liu et al. [19] found that when visualizing hurricane tracks, a selected number of potential tracks and a targeted annotation can help participants estimate storm damage. However, implications of uncertainty visualizations go beyond spatial data. Kale et al. [15] investigated different static and non-static uncertainty visualizations for the detection of historical trends in job data and found non-static visualizations to require less data for effective trend detection. In a weather-related decision scenario, Padilla et al. [23] observed that when quantitative and qualitative uncertainty in probability distributions was high, study participants were more likely to choose a smaller fee than to risk a larger penalty. Kale et al. [14] found that visualizing distributions of a team's predicted score without a mean leads to a better assessment of the impact of individual players on the team's performance. Karduni et al. [16] found that uncertainty visualizations lead to fewer changes in participants' beliefs about time-constant linear relationships between two variables.

In a nutshell, most prior research has studied (a) whether showing predictive uncertainty has any effect at all on people's judgment and decision-making, focusing on only one type of uncertainty visualization [20, 32, 33], or (b) which type of visualization is most effective in communicating uncertainty in general, not specifically focusing on the task of time series forecasting [10, 15, 23]. Given the high relevance of forecasting, we argue that it is important to bring these two streams of research together. Hence, in this paper we study the following research questions:

1. Does exposing predictive uncertainty affect how closely users follow the predictions of a time series forecasting model?

2. Does this effect depend on the type of visualization used to represent predictive uncertainty?

To address these questions, we conducted an online experiment on forecasting COVID-19-related hospitalizations with 79 participants. The participants were shown multiple historical time series of the number of occupied hospital beds due to COVID-19 and had to manually forecast the time series three weeks into the future. They were then shown algorithmic forecasts either without uncertainty information (control group) or with one of two different types of uncertainty visualizations (i.e., 95% confidence interval plots or ensemble displays). Finally, they had the chance to adjust their initial forecast. At this point, we measured how much they adjusted their initial forecast after seeing the algorithmic forecast and how closely their final forecast agreed with the algorithmic forecast.

Surprisingly, contrary to our original hypotheses, we found that when users were shown ensemble displays - a type of visualization that has proven to be very effective in communicating uncertainty - they were less likely to follow the predictions of the forecasting model. To explain this empirical finding, we instantiate the cognitive framework for decision-making with visualizations by Padilla et al. [22] for the case of uncertainty in time series forecasting.

The remainder of this paper is structured as follows: In the next section we develop our hypotheses based on theory and empirical findings from the fields of decision-making and visualization. Next, we describe the design of our experiment in detail. In Section 4, we present the results of our data analysis. Subsequently, we discuss potential reasons for our surprising findings. We close with a brief outlook on future research.

2 RELATED WORK & HYPOTHESES DEVELOPMENT

2.1 Uncertainty vs. No Uncertainty

Our first two hypotheses[1] address Research Question 1. The starting point of our argumentation is the observation that generally

*e-mail: dirk.leffrang@upb.de

†e-mail: oliver.mueller@upb.de

[1]We have pre-registered these hypotheses as well as the rest of the study at https://aspredicted.org/ev2fb.pdf

algorithms tend to be superior to humans when it comes to making predictions [8]. According to Grove's et al. meta-analysis [8], algorithms outperform humans in terms of predictive accuracy by a margin of approximately 10%. Hence, there exists strong empirical evidence that decision makers should consider the outputs of algorithms when making forecasts.

Several studies found providing users with outputs in the form of uncertainty in algorithmic predictions is generally beneficial. For example, in laboratory experiments using weather-related road maintenance games, participants made better decisions and took less risks when provided with information about the uncertainty associated with temperature forecasts [12, 26]. Likewise, in an online experiment using a farming game, participants performed better when they were shown uncertainty estimates of weather forecasts in the form of confidence intervals [7]. In a recent experimental study on house price prediction, McGrath et al. [20] found that people align their own predictions more with provided algorithmic predictions when they were shown probability distributions associated with the algorithmic predictions. Fernandes et al. [5] investigated the impact of uncertainty representation on decision-making for catching a bus and found uncertainty visualization leading to improved decision-making in comparison to no uncertainty visualization. Zhou et al. [33] investigated the effect of confidence intervals around performance curves of two water pipe outage prediction models on user trust. They found that confidence intervals increased participants' reported trust when the cognitive load of a task was low, while they decreased trust when the cognitive load was high.

To sum up, the majority of existing empirical studies observe a positive impact of uncertainty visualizations on algorithm-supported decision-making. Thus, we hypothesize that people are more likely to follow predictions of a time series forecasting model when predictive uncertainty is displayed compared to no uncertainty being displayed. More specifically:

Hypothesis 1a: People are more likely to follow the predictions of a time series forecasting model when predictive uncertainty is visualized through a 95% confidence interval plot compared to no visualization of uncertainty (control group).

Hypothesis 1b: People are more likely to follow the predictions of a time series forecasting model when predictive uncertainty is visualized through an ensemble display compared to no visualization of uncertainty (control group).

2.2 Ensemble Displays vs. Confidence Interval Plot

The third hypothesis addresses Research Question 2. Prior research on uncertainty visualization has developed and tested a variety of graphical annotations and visual encodings for representing uncertainty [21]. In the context of time series forecasting, the 95% confidence interval plot is arguably one of the most common ways of visualizing uncertainty over time. However, as summarized by Padilla et al. [21], prior research has shown that people often misinterpret this type of plot. A main reason is that people have general difficulties interpreting probabilities. As a consequence, they may substitute probabilistic uncertainty information for deterministic information (e.g., in the context of temperature forecasting they often think that the upper and lower bounds represent daily highs and lows). People may also think of the boundaries of confidence intervals as categorical instead of continuous information (e.g., is a point inside our outside of the interval?). They may also assume that all observations within the interval have the same probability [11]. Students, as well as researchers, struggle to understand confidence intervals correctly in experiments [3, 9]. Kim et al. [17] presented sample-based and distribution representations to their study participants, but observed that participants did not necessarily update their prior opinions about statistical relationships rationally and in accordance with the presented statistics. This suggests that both the disagreement between machine and human prediction as well as the adjustment of human predictions based on statistical visualizations are of interest.

An alternative to confidence interval plots are so-called ensemble displays. Originally developed for the purpose of storm forecasting, ensemble displays are generated by repeatedly permuting model parameters and simultaneously drawing all resulting predictions as lines on the same plot [18]. Ensemble displays are a frequency-based way of visualizing uncertainty. For most people, it is easier to interpret frequencies, which they are used to encounter in the real world, than abstract probabilities [6]. In addition, people are not necessarily overwhelmed when confronted with multiple metrics instead of a single aggregated metric. They can naturally aggregate frequencies and calculate summary statistics based on noisy local characteristics [2]. On the contrary, people can interpret different courses displayed as complementary rather than contrary. As a result, taking the average of several noisy metrics can even be more precise than a single metric on its own [1].

In laboratory experiments on forecasting storm paths, ensemble displays outperformed other visualization techniques in terms of interpretability and accuracy [24, 27]. Tak et al. [29] examined differences between ensemble displays and confidence interval plots. Participants were presented with a diagram of predictions of boundaries between different layers of the earth and asked to which layer a particular point on the graph belonged. They found that ensemble displays can visualize normal distributions and uncertainty well. Overall, in most of the related work, frequency-based ensemble displays lead to better decisions compared to probability-based plots (e.g., confidence intervals). Hence, we formulate Hypothesis 2 as follows:

Hypothesis 2: People are more likely to follow the predictions of a time series forecasting model when predictive uncertainty is visualized through an ensemble display compared to a 95% confidence interval plot.

3 METHODS

3.1 Materials & Procedures

The domain of our experiment was the Coronavirus (COVID-19) pandemic. We presented participants with time series plots of the number of hospital beds occupied by COVID-19 patients. In an introduction, we explained participants how to read time series plots and uncertainty information in time series forecasts. Instructions were simple and short in order to reduce cognitive load [3]. The participants' task was to forecast how the time series will develop in the future. More specifically, they had to make forecasts for three points in time (i.e., in 7 days, in 14 days, in 21 days) for nine different countries. Names of countries were hidden and their order was randomized. For each of the nine countries, the task comprised the following two steps:

1. We showed participants the historical development of the number of occupied hospital beds due to the COVID-19 pandemic. Figure 2 shows an example of such a time series. Based on this information, they had to make forecasts for t+7, t+14, and t+21 days.

2. After submitting their own forecasts, we provided an algorithmic forecast from a Bayesian structural time series model. This forecast could contain different information about the uncertainty of the predictions. After seeing the algorithmic forecasts, participants had the chance to adjust their own forecasts.

Figure 1 shows an example of the user interface in the second step. As participants change their predictions in the first or second step, the point predictions are interactively adjusted in the diagram. In the first step, the point predictions for t+7, t+14 and t+21 are anchored at the value of t. User inputs from the first step form the anchors for the point predictions in the second step.

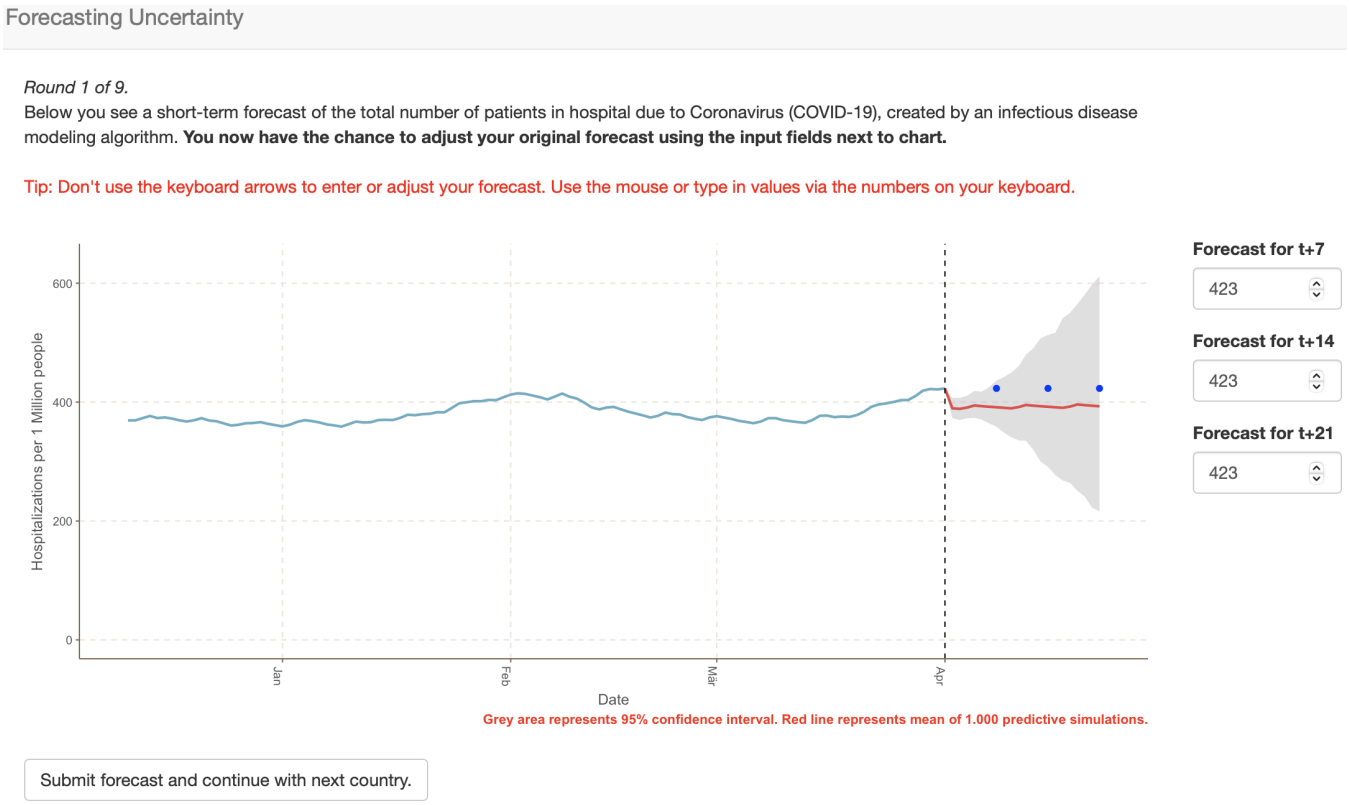

Figure 1: Example of a user interface for a CI in the second step.

In order for participants to become familiar with the procedures and materials, we defined the first three tasks as warm-up tasks and excluded them from the subsequent data analysis.

3.2 Participants & Conditions

We invited 126 first-year Bachelor students (Management Information Systems) to participate in the experiment. They could choose to participate in the online experiment between June 8th and June 14th, 2021. Students received a performance incentive in the form of exam bonus points for their participation. Not participating in the experiment did not results in any penalty. 84 of the invited students participated in the experiment. If a participant's data was incomplete, they were excluded. Furthermore, if a participant completed the experiment too fast (i.e., less than 90 seconds for all 9 tasks) or too slow (i.e., more than two standard deviations above the mean participation time), their data was excluded. This prevented participants from just "clicking through" the experiment. Five of the participants were excluded due to these criteria.

There were three conditions in the experiment, varying how predictive uncertainty of the algorithmic time series forecast was visualized. For each task, participants were randomly assigned to one of the three conditions. Hence, all participants were exposed to all three conditions over the course of the nine tasks.

A point estimate with no uncertainty information is the first condition and serves as control group. Figure 3 shows an example.

The treatment for the second condition was a confidence interval (CI) plot, showing the most probable prediction as well as a 95% CI around it represented by a shaded area. We derived the CI from a posterior predictive distribution of a Bayesian structural time series model.[2] Figure 4 depicts an example of the CI plot.

[2]Since the predictive model used in this study was a Bayesian model, the uncertainty intervals were actually credibility intervals. Yet, our participants were more familiar with confidence intervals than credibility intervals due to their previous statistics courses. Furthermore, the visualization of credible intervals is similar to the one for confidence intervals. Therefore, we labeled credible intervals as confidence intervals in our experiment.

Our third condition was an ensemble display. We generated the lines of the ensemble by sampling from the posterior predictive distribution of the Bayesian structural time series model. In our experiment, we displayed twenty random draws from the model. Figure 5 shows an example.

3.3 Model Specification

Inspired by the experiments of Poursabzi-Sangdeh et al. [25] and McGrath et al. [20], we operationalized 'willingness to follow predictions' by two variables. First, the difference between the first and second prediction served as a measurement for the adjustment due to seeing the algorithmic prediction. Adjustment captures the extent to which participants have updated their own prior belief after seeing the algorithmic prediction and uncertainty specification [20]. Second, the difference between a participant's final prediction and the algorithmic prediction served as a measurement for the disagreement between human and algorithm. If there are large discrepancies between the algorithmic predictions and the participants' final predictions, then the participants obviously did not attach a lot of importance to the model's forecasts in their own decision.

We used Bayesian hierarchical and multi-level models with varying intercepts for condition, participant, and task. As dependent variable Y we used either adjustment[3] or disagreement. We defined each of the nine countries as a task. An observation i was defined

[3]As for some participants the adjustment variable was zero, we added a small amount (0.001) to adjustment values to avoid errors due to the undefined logarithm of zero.

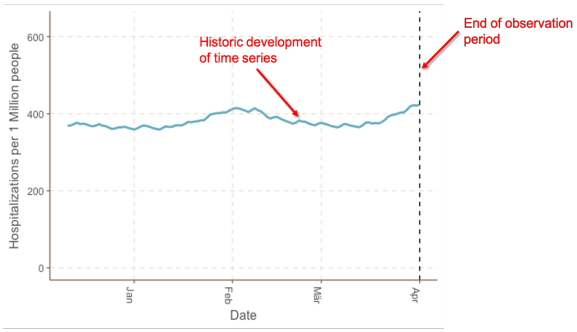

Figure 2: Historical development of a COVID-19 hospitalization time series.

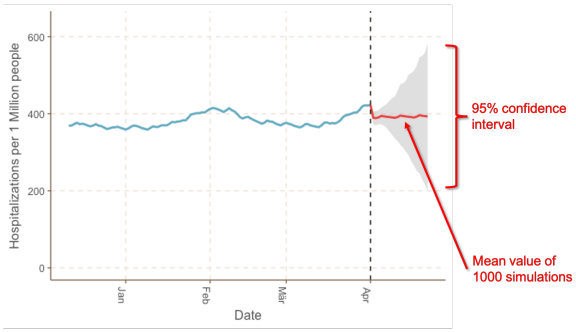

Figure 4: 95% CI plot. Mean value of 1000 simulations in red, 95% confidence interval in gray.

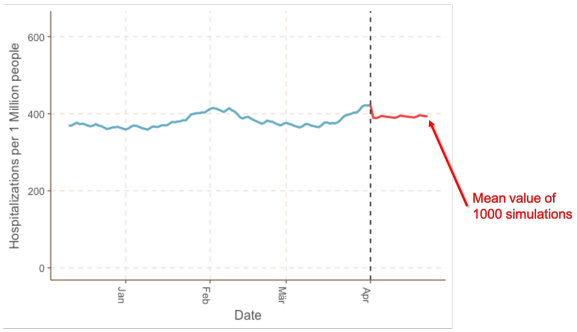

Figure 3: Line plot of point estimates without uncertainty.

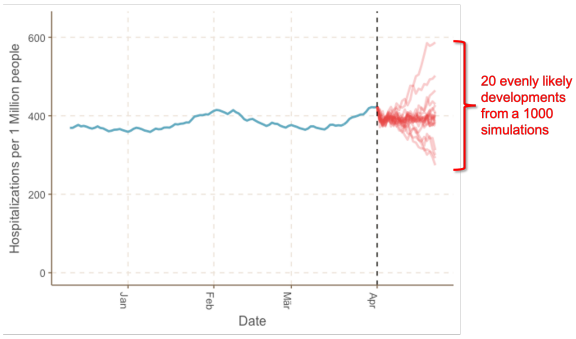

Figure 5: Ensemble display for prediction. Most probable prediction in red, other likely results in gray.

as a participant-task combination. For each participant, the order of the tasks and the condition for each task were randomized. This resulted in the following generic model specification:

$$\log(Y_i) = \gamma_{condition[i]} + \pi_{participant[i]} + \tau_{task[i]} + \varepsilon_i$$

In addition to calculating condition mean effects, we also calculated the contrasts between conditions to assess whether there are any significant differences between conditions.

4 Results

4.1 Model-Free Evidence

A total of 474 predictions, that is participant-task combinations, were collected during the experiment. The distributions of the predictions are shown in Figure 6. Eyeballing the violin plots suggests that when participants were shown ensemble displays, they adjusted less after seeing the model's forecasts and disagreed more with the model's forecasts.

We also calculated the absolute error between participants' initial forecasts and the true values (Mean: 31.46, Median: 17.70) and between the algorithmic forecasts and the true values (Mean: 33.22, Median: 10.86). The clearly lower median error suggests that participants should have followed the model in order to minimize overall forecasting error (the approximately even means resulted from large outliers).

In summary, the model-free results surprisingly indicate a potential negative effect of uncertainty visualization on the willingness to follow a model's predictions; although the model's forecasts were more precise than the manual forecasts in most cases.

4.2 Estimation results

To get a clearer picture of these patterns, we calculated the differences of mean effects between conditions (see Table 1).

Neither for adjustment (0.05 [-0.57, 0.69]) nor for disagreement (-0.01 [-0.25, 0.23]) we found significant differences between CI plots and point estimate plots with no uncertainty. This does not confirm H1a, that people are more likely to follow the predictions of a time series forecasting model when predictive uncertainty is visualized through a 95% confidence interval plot compared to no visualization of uncertainty (control group).

Regarding the contrasts between ensemble displays and plots of point estimates with no uncertainty, we found significant differences for both adjustment (-0.79*** [-1.47, -0.10]) and disagreement (0.37*** [0.11, 0.62]). However, the observed effects go into the opposite direction of our hypothesis H1b that people are more likely to follow the predictions of a time series forecasting model when predic-

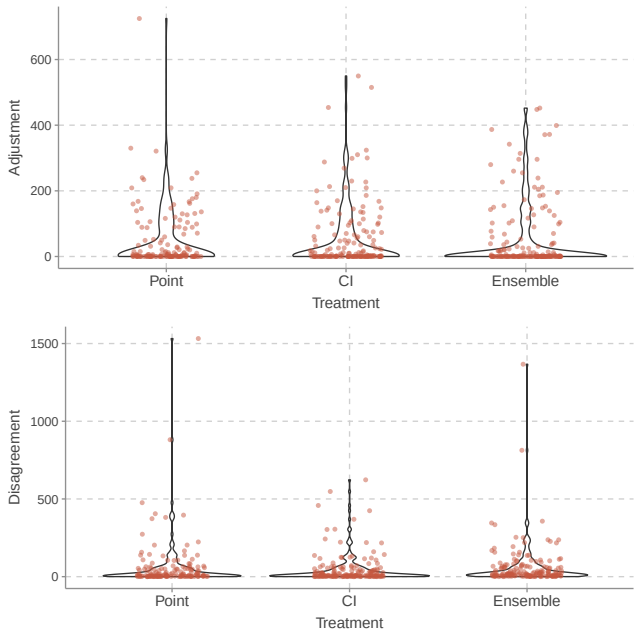

Figure 6: Violin plots for ln-transformed values of adjustment (top) and disagreement (bottom) by conditions.

Table 1: Condition mean effects and contrasts of adjustment and disagreement along conditions.

Adjustment				
Condition mean effect		Estimate		95% CI
	Point	0.26		[-1.72, 2.28]
	CI	0.32		[-1.64, 2.34]
	Ensemble	-0.53	*	[-2.49, 1.46]
Contrasts between conditions		Difference		95% CI
H1a:	CI - Point	0.05		[-0.57, 0.69]
H1b:	Ensemble - Point	-0.79	***	[-1.47, -0.10]
H2:	Ensemble - CI	-0.86	***	[-1.73, -0.16]
Disagreement				
Condition mean effect		Estimate		95% CI
	Point	2.48	***	[1.37, 3.59]
	CI	2.47	***	[1.36, 3.57]
	Ensemble	2.84	***	[1.74, 3.95]
Contrasts between conditions		Difference		95% CI
H1a:	CI - Point	-0.01		[-0.25, 0.23]
H1b:	Ensemble - Point	0.37	***	[0.11, 0.62]
H2:	Ensemble - CI	0.38	***	[0.12, 0.62]

Notes: Significance levels: * 90%, ** 95%, *** 99% CI does not contain 0.

tive uncertainty is visualized through an ensemble display compared to no visualization of uncertainty (control group). Because ensemble displays lead to less adjustment and more disagreement, we can infer that in ensemble display conditions, participants followed the forecasting model to a lesser extent than in the no uncertainty condition.

Finally, we tested H2 that people are more likely to follow the predictions of a time series forecasting model when predictive uncertainty is visualized through an ensemble display compared to a 95% confidence interval plot. We found significant differences in both adjustment (-0.86*** [-1.73, -0.16]) and disagreement (0.38*** [0.12, 0.62]) between the ensemble display and CI conditions. Again, the direction of the effects goes into the opposite direction of our hypothesis. In the ensemble display conditions, participants followed the model less than in the CI condition.

5 Discussion

Uncertainty is an important factor for making informed decisions. This paper extends the existing body of knowledge by examining the effect of visualizing predictive uncertainty on users' willingness to follow the predictions of a time series forecasting model. Surprisingly, our findings contradict most prior empirical results and also our own hypotheses.

First, we could not confirm that participants are more likely to follow predictions of a time series forecasting model when predictive uncertainty is visualized through a 95% CI plot compared to no visualization of uncertainty (H1a).

Second, we could also not confirm that ensemble displays are more likely to result in participants following the algorithm compared to point estimates with no uncertainty visualizations (H1b). On the contrary, we observed the opposite of what we initially hypothesized.

Third, we were also unable to confirm that participants are more likely to follow predictions of a time series forecasting model when predictive uncertainty is visualized through an ensemble display compared to a 95% CI plot. Again, we observed the opposite of our initial expectations. Our empirical results suggest that more salient visualizations of uncertainty lead to decreased willingness to follow algorithmic forecasts.

Most studies on uncertainty visualization only test effects rather than provide explanations for the observed effects [10]. To this end, in the following we synthesize insights from different theoretical research streams on decision-making and uncertainty visualization. Padilla et al. [22] presented a cognitive framework for visual decision-making. Figure 8 shows an extract of the framework. According to the framework, people use a graph to find an answer to a conceptual question. An instantiated graph schema is the interpretation of a mentally constructed visual description of a raw visual array. During the message assembly process people try to extract a conceptual message from the instantiated graph schema; a process which is influenced by the conceptual question at hand. Conceptual messages subsequently influence decision-making processes, which in turn inform behavior.

The original model addresses decision-making in general. In the following, we instantiate the original framework by focusing on the special case of visualizing predictive uncertainty. An effective uncertainty visualization is an example of an instantiated graph schema. If an uncertainty visualization is effective, and aligns with the question at hand, the message assembly process results in a conceptual message expressing a higher degree of uncertainty awareness. To explain how the message of high uncertainty informs participants' decision-making and behavior, we draw on psychological research about advice giving and taking. Specifically, prior research, as summarized by Bonaccio et al. [4], found that an advice-seeker is less likely to follow advice if he or she has the impression that the advice-giver is uncertain. Therefore, greater uncertainty awareness should lead to less advice utilization. To sum up, more effective uncertainty visualizations lead to higher uncertainty awareness (left hand side of Figure 8), which lead to reduced advice utilization (right hand side of Figure 8).

Applying this model to our empirical results suggests that point and CI plots were similar in their effectiveness of uncertainty visualization. A reason could be that for both plots participants focused on the mean values and discarded the actual confidence intervals. Ensemble displays, in contrast, seemed to be a more effective way of uncertainty visualization. They show multiple possible future

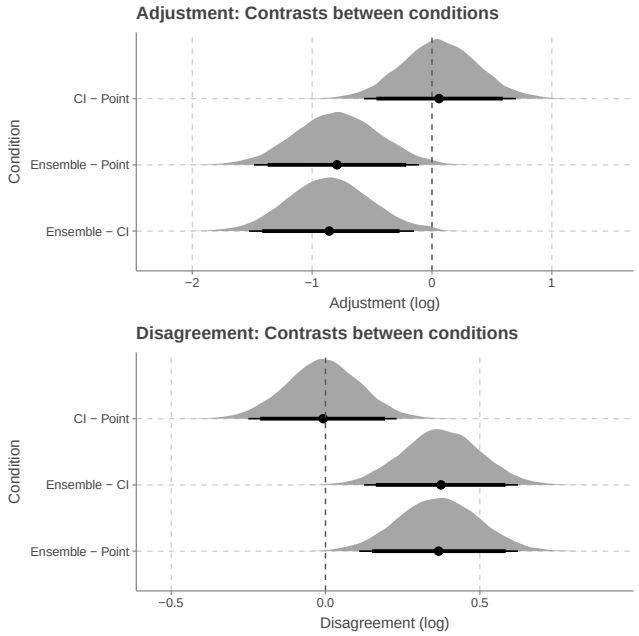

Figure 7: Contrasts between conditions for adjustment (top) and disagreement (bottom).

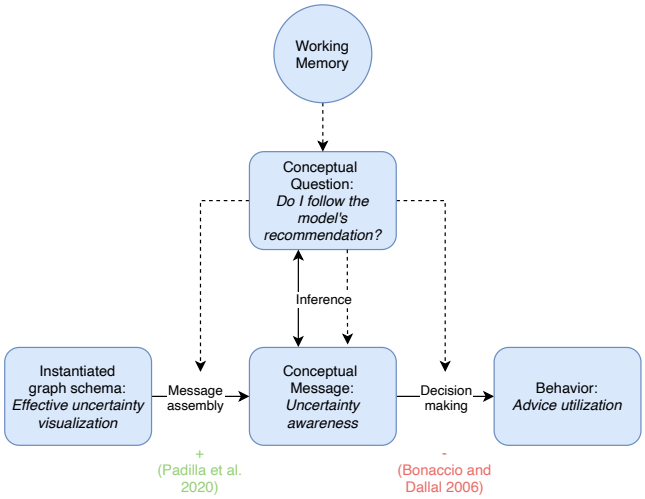

Figure 8: Instantiated extract from the model of visualization decision-making adapted from Padilla et al. [21].

scenarios and do not explicitly show their mean value, which may lead to more uncertainty awareness and, in turn, less advice utilization. In other words: Becoming aware of the uncertainty of the algorithmic advice giver, advice seekers probably decided to not follow the advice and stay with their initial manual predictions.

Our findings have several implications for practice and research. If visualization of uncertainty has an impact on the decision-making process, it can be exploited strategically. This is a double-edged sword. On the one hand, there are good arguments for following algorithmic predictions, as they tend to be more accurate than manual predictions [8]. On the other hand, from an ethical point of view, people should not blindly follow algorithms [28]. Visualization of uncertainty is therefore a tool which should be used purposefully and with caution.

6 Limitations & Outlook

As our experimental results surprisingly contradict prior empirical studies, we argue for more studies on the effect of visualizing predictive uncertainty on decision maker's willingness to follow algorithmic advice. Most importantly, we need a better understanding of the causal mechanisms between different types of uncertainty visualizations and different types of user tasks. The framework for visual decision-making [21] is a promising theoretical foundation for this, but needs to be instantiated for more concrete domains and tasks.

In addition, in future work our experiments should be replicated and extended with regards to other tasks and visualizations. For example, uncertainty range width, the coloring of graphs, and labeling of graphs may have an influence on the perceived predictive uncertainty [29].

Of course, our study is not without limitations. First, the reward of participants in our experiment did not depend on their performance in the experiment. We communicated that when participants completed all tasks conscientiously, they would receive their exam bonus points. But we intentionally only loosely defined "conscientiously" as depending on certain proximity measures to prevent opportunistic behavior. Future research may test whether participants' incentives influence the degree to which they follow algorithmic predictions. In addition, Management Information Systems students tend to be data-savvy and open-minded towards technology. Other user groups may have different background knowledge of and attitudes towards algorithmic recommendations and uncertainty visualizations. Hence, future research should investigate whether user groups with different knowledge of and attitudes towards algorithms may be influenced differently. In addition, strong opinions of participants about COVID-19 may have biased the results. However, randomization should have equalized such non-structural differences.

In addition, further research is needed to address the question why exactly participants adjusted their predictions. The reason could be an actual increased awareness of uncertainty, as argued above, or an unconscious reaction to a visual stimuli (e.g., anchoring on the mean line or decreased perceptual accuracy without a mean line). For example, point estimates and CI may have facilitated perception of algorithmic predictions due to their lower visual complexity, whereas random posterior draws may have confused participants. An examination of incentive-based decision-making, as well as qualitative criteria, can also extend the findings of this study.

We focused on static visualization techniques because they are most universally applicable across all media and, based on the media examples we identified, are most common in the use case under consideration. Comparisons with other visualization techniques, particularly dynamic ones such as Hypothetical Outcome Plots, represent another promising area of research.

Furthermore, prior research has documented learning effects in visualization experiments [15]. We tackled this potential threat to validity by randomization and a relatively small amount of tasks. Future research should examine how different visualizations influence users' decision-making with regards to the ordering of and time spend with visualizations, as well as the familiarity with a subject area or task.

Finally, the anchoring of initial values on the value in t may have distorted the results of the predictions for t+7, t+14 and t+21 before the algorithm was shown [31]. Therefore, future research might investigate whether different anchors have an impact on people's willingness to follow algorithmic forecasts.

References

[1] G. A. Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3):122–131, 2011. doi: 10.1016/j.tics.2011.01.003

[2] G. A. Alvarez and A. Oliva. The representation of simple ensemble visual features outside the focus of attention: Research article. *Psychological Science*, 19(4):392–398, 2008. doi: 10.1111/j.1467-9280.2008.02098.x

[3] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4):389–396, 2005. doi: 10.1037/1082-989X.10.4.389

[4] S. Bonaccio and R. S. Dalal. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2):127–151, 2006. doi: 10.1016/j.obhdp.2006.07.001

[5] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In *Proceedings of the 2018 Conference on Human Factors in Computing Systems (CHI'18)*, pp. 1–12. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173718

[6] G. Gigerenzer and U. Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995. doi: 10.1037/0033-295X.102.4.684

[7] M. Greis, P. E. Agroudy, H. Schuff, T. Machulla, and A. Schmidt. Decision-making under uncertainty: How the amount of presented uncertainty influences user behavior. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI'16)*, pp. 1–4. ACM, New York, NY, USA, 2016. doi: 10.1145/2971485.2971535

[8] W. M. Grove, D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson. Clinical versus mechanical prediction: A meta-analysis. *Psychological assessment*, 12(1):19–30, 2000. doi: 10.1037/1040-3590.12.1.19

[9] R. Hoekstra, R. D. Morey, J. N. Rouder, and E. J. Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review*, 21(5):1157–1164, 2014. doi: 10.3758/s13423-013-0572-3

[10] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, 2019. doi: 10.1109/TVCG.2018.2864889

[11] H. Ibrekk and M. G. Morgan. Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, 7(4):519–529, 1987. doi: 10.1111/j.1539-6924.1987.tb00488.x

[12] S. L. Joslyn and J. E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, 18(1):126–140, 2012. doi: 10.1037/a0025185

[13] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY, USA, 2011.

[14] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2021. doi: 10.1109/TVCG.2020.3030335

[15] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):892–902, 2019. doi: 10.1109/TVCG.2018.2864909

[16] A. Karduni, D. Markant, R. Wesslen, and W. Dou. A Bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):978–988, 2021. doi: 10.1109/TVCG.2020.3029412

[17] Y. S. Kim, L. A. Walls, P. Krafft, and J. Hullman. A Bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI'19)*, pp. 1–14. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300912

[18] L. Liu, A. P. Boone, I. T. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2165–2178, 2017. doi: 10.1109/TVCG.2016.2607204

[19] L. Liu, L. Padilla, S. H. Creem-Regehr, and D. H. House. Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):882–891, 2019. doi: 10.1109/TVCG.2018.2865193

[20] S. McGrath, P. Mehta, A. Zytek, I. Lage, and H. Lakkaraju. When does uncertainty matter?: Understanding the impact of predictive uncertainty in ML assisted decision making. *arXiv*, 2020.

[21] L. Padilla, M. Kay, and J. Hullman. Uncertainty visualizations. *To appear in Handbook of Computational Statistics and Data Science*, 2020.

[22] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, 3(1), 2018. doi: 10.1186/s41235-018-0120-9

[23] L. M. Padilla, M. Powell, M. Kay, and J. Hullman. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, 11(579267), 2021. doi: 10.3389/fpsyg.2020.579267

[24] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive research: principles and implications*, 2(1):1–16, 2017. doi: 10.1186/s41235-017-0076-1

[25] F. Poursabzi-Sangdeh, D. G. Goldstein, and J. M. Hofman. Manipulating and measuring model interpretability. In *Proceedings of the 2021 Conference on Human Factors in Computing Systems (CHI'21)*, pp. 1–52. ACM, New York, NY, USA, 2021. doi: 10.1145/3411764.3445315

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

[27] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition and Computation*, 16(2):154–172, 2016. doi: 10.1080/13875868.2015.1137577

[28] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the 2015 ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*, pp. 141–148. ACM/IEEE, New York, NY, US, 2015. doi: 10.1145/2696454.2696497

[29] S. Tak, A. Toet, and J. Van Erp. The perception of visual uncertainty representation by non-experts. *IEEE Transactions on Visualization and Computer Graphics*, 20(6):935–943, 2014. doi: 10.1109/TVCG.2013.247

[30] C. Tannert, H. Elvers, and B. Jandrig. The ethics of uncertainty: In the light of possible dangers, research becomes a moral duty. *EMBO reports*, 8(10):892–896, 2007. doi: 10.1038/sj.embor.7401072

[31] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124

[32] Y. Zhang, Q. Vera Liao, and R. K. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*'20)*, pp. 295–305. ACM, New York, NY, USA, 2020. doi: 10.1145/3351095.3372852

[33] J. Zhou, S. Z. Arshad, S. Luo, and F. Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *2017 Conference on Human-Computer Interaction (INTERACT'2017)*, pp. 23–39. Springer International Publishing, Bombay, India, 2017. doi: 10.1007/978-3-319-68059-0_2