International Journal of Forecasting 41 (2025) 689–701

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

# Improving out-of-population prediction: The complementary effects of model assistance and judgmental bootstrapping

Mathew D. Hardy [a,*], Sam Zhang [b], Jessica Hullman [c], Jake M. Hofman [d], Daniel G. Goldstein [d]

[a] Department of Psychology, Princeton University, United States of America
[b] Department of Applied Mathematics, University of Colorado Boulder, United States of America
[c] Department of Computer Science, Northwestern University, United States of America
[d] Microsoft Research, United States of America

ARTICLE INFO

*Article history:*
Dataset link: https://github.com/mdahardy/judgmental-bootstrapping

*Keywords:*
Judgemental forecasting
Bootstrapping
Adjusting forecasts
Decision making
Time Series
Model Selection

ABSTRACT

We propose and test a method for out-of-population prediction termed model-assisted judgmental bootstrapping, which leverages a predictive model from one domain combined with expert judgment to generate training data and subsequently a predictive model for a new domain. In a preregistered experiment ($N$=1440), we assessed the predictive accuracy of this method in increasingly challenging environments. We also analyzed the individual contributions of two techniques that underlie the method: model-assisted estimation and judgmental bootstrapping. Our findings revealed that both techniques significantly improved predictive accuracy. Furthermore, their impacts were complementary: model-assisted estimation provided the largest accuracy gains in the least demanding environment, while judgmental bootstrapping did so in the most challenging environment. Our results suggest that model-assisted judgmental bootstrapping is a promising technique for creating predictive models in domains in which outcome data are not available.



## 1. Introduction

In early 2020, the world watched in alarm as Italy grappled with a rapid surge of a mysterious new respiratory virus. The streets of Milan and Rome, normally bustling with life, were empty and quiet as the nation went into lockdown. As countries around the world braced for similar outbreaks, epidemiologists faced an urgent challenge: to forecast the trajectory of this virus in their countries using data from Italy, despite sometimes large differences in demographics, healthcare infrastructure, and cultural norms.

Making predictions about the local spread of Covid-19 became a pressing concern for researchers and governments around the world (Ferguson et al., 2020; Remuzzi & Remuzzi, 2020). One reason the task was so difficult was that researchers had to make *out-of-population* predictions. These are predictions using models and data from one domain (e.g., Italy) for a new, untested domain (e.g., the US) where no predictive model or outcome data are available. This is in contrast to out-of-sample (but in-population) prediction, where models are tested on held-out or new cases from the same domain Katsikopoulos, Simsek, Buckmann, and Gigerenzer (2021) and Todd and Gigerenzer (2012).

While difficult, making predictions about a new and unmodeled domain arises in many forecasting tasks. In

* Correspondence to: Princeton University, Department of Psychology, Peretsman Scully Hall, Princeton, NJ 08540, United States of America.
*E-mail address:* mdhardy@princeton.edu (M.D. Hardy).

The numerical results presented in the manuscrit were reproduced by the Editor-in-Chief on the 12th of September 2024.

https://doi.org/10.1016/j.ijforecast.2024.07.002

business, out-of-sample prediction could involve forecasting the monthly sales of a new franchise store in an established market with many existing locations. In contrast, out-of-population prediction could entail forecasting the sales of the first-ever store in an entirely new market (e.g., a dissimilar country). In climate science, out-of-sample predictions include forecasts about Atlantic hurricanes based on decades of recorded events. Conversely, a meteorologist would make an out-of-population prediction when forecasting hurricane trajectories in California using models developed based on the behavior of Atlantic hurricanes. In economics, an out-of-sample prediction might involve estimating GDP growth for a country based on its historical data, while an out-of-population prediction could involve using a model trained on one country's economic indicators to forecast growth for another country that lacks high-quality economic indicator data.

How can one train a model to make predictions in a new domain where outcome data are not available? In this paper, we propose and test a method termed *model-assisted judgmental bootstrapping*. The approach is rooted in two ideas from the forecasting literature. The first is the concept of model assistance during the process of expert elicitation (Lawrence, Goodwin, O'Connor, & Önkal, 2006), and the second is a technique known as judgmental bootstrapping (Armstrong, 2001). Model assistance aids in eliciting accurate estimates, and judgmental bootstrapping uses these estimates to train a predictive model. Through a large randomized experiment, we find that both model assistance and judgmental bootstrapping improve accuracy in out-of-population prediction. Importantly, we find that their impacts are complementary: model assistance proved most beneficial in less challenging environments, while judgmental bootstrapping did so in more challenging ones.

### 1.1. Terminology

In model-assisted judgmental bootstrapping, we assume there is a person (the *expert*), who is able to obtain predictions from a model trained in one domain (termed the *old domain*) but aims to train a model for predictions in a different domain (referred to as the *new domain*). It is important to note that the new domain may be similar to the old domain. However, there are no outcome data for the new domain, so the expert cannot train a model to make predictions.

A *case* is defined as a set of predictor values (for instance, economic indicator values of a country). The act of *calling a model* involves *passing* (i.e., providing as input) a case to a predictive model to obtain a prediction. Lastly, a *target* case represents a specific instance in the new domain for which a prediction is desired. With these terms and assumptions in hand, we next review the areas of the forecasting literature in which the technique is rooted.

### 1.2. Roots of the approach

One of the two foundations of the method is judgmental bootstrapping, a technique which, according to Dawes and Corrigan (1974), was first proposed by Yntema and Torgerson (1961). In judgmental bootstrapping, experts are presented with cases and then use their domain knowledge to provide educated estimates of outcomes. Once these estimates are compiled, a model is developed to predict the expert's estimates. Subsequently, this model (rather than the expert) is used to make predictions on new cases.

Why opt for a model based on experts' estimates instead of directly using the experts' estimates? This is because models generated through judgmental bootstrapping tend to match or outperform the estimates on which they are trained. As early as 1979, Dawes characterized judgmental bootstrapping as "pervasive" (p. 575) and cited only one exception to its efficacy (Dawes, 1979). Later, Armstrong 2001 conducted a review and observed that judgmental bootstrapping consistently yields modest gains in predictive accuracy. Various studies, such as those by Camerer (1981) and Karelaia and Hogarth (2008), have investigated the conditions favorable for this technique. Because it uses experts' estimates instead of outcome data, judgmental bootstrapping aligns with our objective of crafting a model when outcome data are not available.

The second foundation of our approach is model-assisted estimation, which is informed by the judgmental forecasting literature (Lawrence et al., 2006), especially studies on judgmental adjustments of statistical forecasts (Dietvorst, Simmons, & Massey, 2018; Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Goodwin & Fildes, 1999; Lim & O'Connor, 1995; Sanders & Ritzman, 2001) and advice (Harvey, Harries, & Fischer, 2000). We conceptualize model-assisted estimation as a three-step process. Presuming there is an expert who is able to call an old-domain model, in the first phase the expert uses their judgment to construct a case to provide to the old-domain model that is expected to yield an informative prediction. In the subsequent step, the old-domain model produces a prediction for this case. In the last step, the expert reviews the old-domain model's prediction and records their best estimate for the new domain.

Drawing on the rich literature on judgmental bootstrapping and judgmental adjustment (an important step in model-assisted estimation) this work's contribution lies in (i) specifying the process for model-assisted judgmental bootstrapping, (ii) conducting a highly powered, preregistered experiment that estimates the effects of model-assisted estimation and judgmental bootstrapping on predictive accuracy, and (iii) examining the contribution of both model-assisted estimation and judgmental bootstrapping in increasingly challenging task environments. We begin by detailing each step of the full process and provide an example.

### 1.3. Model-assisted judgmental bootstrapping

As illustrated in Fig. 1, when presented with a target case in a new domain, the expert first identifies an

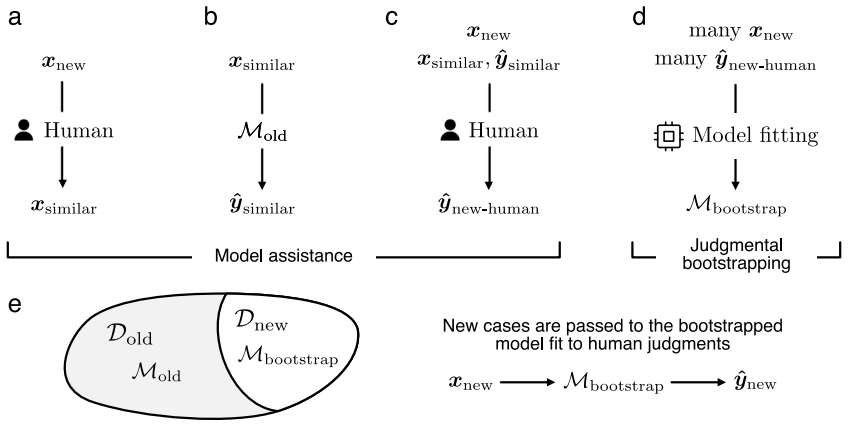


**Fig. 1.** Given an old domain $\mathcal{D}_{old}$ and a new domain $\mathcal{D}_{new}$, model-assisted judgmental bootstrapping is a five-step process. (a) Based on the target case in the new domain $\boldsymbol{x}_{new}$, an expert identifies a set of predictor values $\boldsymbol{x}_{similar}$ to input into the old-domain model $\mathcal{M}_{old}$ in order to obtain a prediction that the expert can consult later. (b) The old-domain model creates a prediction for $\boldsymbol{x}_{similar}$ designated $\hat{\boldsymbol{y}}_{similar}$. (c) The expert reviews $\hat{\boldsymbol{y}}_{similar}$ in light of $\boldsymbol{x}_{similar}$ and $\boldsymbol{x}_{new}$ and records their best estimate for the target case in the new domain, which is termed $\hat{\boldsymbol{y}}_{new\text{-}human}$. (d) When a sufficient number of pairs of expert estimates and new cases are in hand, they are used to train a judgmental bootstrapping model $\mathcal{M}_{bootstrap}$ to predict the expert's estimates. (e) The bootstrapping model is ready to forecast new cases in the new domain.

old domain model and case to pass to this model. The expert does so to obtain useful predictions for making an estimate for the target case. The expert thus bases this selection on both the target case and the old-domain models to which the expert has access.

Passing the selected case to the old-domain model results in an old-domain prediction. The expert reviews this old-domain prediction and records their best estimate for the target case in the new domain. Once an adequate number of these expert estimates have been gathered, they are channeled into the judgmental bootstrapping process. Specifically, a model is trained to predict the experts' estimates based on the cases from the new domain. Modelers can use well-known techniques from statistics and machine learning to train and evaluate this model. Once trained, the new model—rather than the expert—can be used to make predictions in the new domain. Note that the final bootstrapping model depends only on participants' estimates and may even be based on different features from those used in the old-domain model.

To illustrate, consider an ice-cream franchise with multiple outlets in New England in the US. This franchise is preparing to open its first international outlet in a distant country and wants to forecast its monthly sales for the initial 12 months. The company has constructed a model that accurately predicts monthly sales for its New England outlets and has been using this model to project sales for the first 12 months at various new branches in New England. However, this New England-based (i.e., old-domain) model relies on 10 predictors. Out of these, only three are available in the new domain. Moreover, there is an additional predictor that is useful in the new domain but is not present in the old one. A challenge emerges as the expert wishes to utilize the insights from the New England model for forecasts, but the predictors between the New England and new domains are not fully aligned, and passing new-domain cases to the old domain model is not feasible. To circumvent this issue, the expert exercises judgment to construct similar cases (step a in Fig. 1) that can be passed to the New England model (step b). For instance, predictor values might be adopted from a New England store that the expert believes would yield an informative prediction for the new location. The expert will later bring to bear all that they know about the new location in order to adjust this prediction (step c) before sending it into the judgmental bootstrapping process (step d). The result is a model that makes predictions in the new domain using only the new-domain predictors (step e).

## 2. Task environments

The task in our experiment involved predicting values in a time series. Time series forecasts are instrumental for planning across various domains, from epidemiology to business, and have been extensively studied in the context of judgmental forecasting and adjustment (e.g., Goodwin & Wright, 1993; Harvey, 1988). Typically, in time series forecasting tasks, participants are provided with a series of past observations to help predict upcoming ones. However, as our interest lies in predictions made in the absence of outcome data, we asked participants to forecast values at multiple future points without referencing past observations.

In the experiment, individuals from an online participant pool predicted the average monthly high temperatures in global cities. In one of the randomly assigned conditions, this is guided by an old-domain model that has only been trained on a sample of major US cities. Before making their predictions, participants select cases for which they wish to view predictions from the US-based model. Specifically, they can ask to see predictions for one of the US cities in the training data. The participants can then consult these predictions while making their estimates for the target city. As our defined out-of-population prediction task presupposes the unavailability of outcome data, no feedback was given to the participants.

In previous sections, we highlighted that model-assisted judgmental bootstrapping involves experts and is intended for out-of-population scenarios where outcome data are not available. Given this, it might appear odd that we are conducting an experiment with a non-expert population about climate, a domain rich with outcome data. Despite these limitations, we chose this participant population and domain due to certain advantages. First, there are an abundance of ground truth data concerning weather and climate, allowing us to assess the accuracy of both participants' and models' estimates. Second, by varying the locations of the target cities we can vary the difference between the best-possible old-domain predictions and the ground-truth values for the target cities. Due to spatial autocorrelation, geographically distant locations tend to have more distinct climates than proximate ones (Di Cecco & Gouhier, 2018). Target cities that are chosen to be close to the old domain should generally require less adjustment. More adjustment would be required when the old-domain model is trained in the Northern hemisphere and the target cities are in the Southern hemisphere, owing to seasonal reversal (by which winter in the Southern hemisphere is summer in the Northern hemisphere and vice versa). Third, *also* by varying the locations of the cities, we can approximate varying the relative expertise of the participants. People are likely to be relatively more knowledgeable about average high temperatures in cities within their country of residence. This could arise because of spatial autocorrelation (cities in an individual's country of residence are generally closer and therefore more similar in temperature to their own city), local travel experience, or the transmission of climate information through domestic media and social networks.

Drawing parallels to Hogarth, Lejarraga, and Soyer (2015), who gained insights by investigating *kind* (straightforward) versus *wicked* (misleading) learning environments, we generated Kind, Challenging, and Wicked task environments, as detailed in Table 1. Given that the US-based participants could request predictions from a model trained on US cities, we assumed that making predictions for other US cities would fall into the Kind task environment. The best-possible predictions from the US model should be relatively close to the ground-truth values of unmodeled US cities. As we show below, well-chosen predictions from the US model require very little adjustment for other US cities. Moreover, participants in the US should have relatively more expertise about the weather in US cities than about cities outside the US. At the other extreme, we imagined that predicting temperatures for cities in the Southern hemisphere based on a model trained on US cities fit in the Wicked domain. Not only would participants be less knowledgeable about weather in these cities, but the best possible predictions from the US model would be inaccurate or even misleading due to distance and seasonal reversal between the Northern and Southern hemispheres. We also suspected that some participants would be unaware of seasonal reversal and fail to account for it in their adjustments. Between these two extremes, we constructed a Challenging domain of cities outside the US that are in the Northern hemisphere. Varying the task environment thus allowed us to vary the accuracy of the candidate models: US model cities are generally better models of Kind-domain cities than they are of Challenging and Wicked cities.

## 3. Experiment

Before running our experiment, all analyses and statistical models were preregistered.[1] We report all planned analyses from our preregistration below.

### 3.1. Participants

We recruited 1440 participants from Amazon's Mechanical Turk participant pool. We determined our sample size through a simulation-based power analysis where we generated bootstrapped datasets by sampling from pilot data to create samples equivalent to our target sample size (more details on the power analysis are given in Appendix A.2). We required that participants have at least 2500 approved HITs (Mechanical Turk tasks), an approval rating of at least 99%, and be located in the United States. Participants were only allowed to complete the experiment once, and participants that completed pilot versions of the experiment were not allowed to participate. Participants received $1.50 for completing the experiment and spent an average of 11.67 min on the task.

### 3.2. Procedure

As introduced above, we used a forecasting task in which participants predicted the average high temperatures for various target cities around the world (see Fig. 2). For each target city, participants provided 12 separate temperature forecasts (one for each month) with predictions given in Fahrenheit and confined to rounded figures between −99° and 999°.[2] We estimated the ground-truth average highs for each city by averaging the observed high temperatures for a weather station near the city center or main city airport between January 1st, 1980 and December 31st, 2020 for each month, using data gathered from the `rnoaa` R package.

As described above, each target city was in one of three task environments: Kind, Challenging, or Wicked. All participants made predictions on one city for each task environment. These three cities were randomly sampled at the start of the experiment from a predefined set of 24 cities, with eight for each domain (see Table 1; city order was randomized for each participant).

Each participant was randomly assigned to one of two conditions. In the Control condition, participants made their predictions independently, using only their own knowledge and judgment. In the Model-Assist condition, participants selected a model city from a pre-defined list to aid them in their predictions before making their predictions for each target city (participants could select a different model city for each trial). The model city represents the similar old-domain case.[3] Model cities consisted

[1] Preregistration available at https://aspredicted.org/qq5yh.pdf.

[2] This asymmetry was due to the text boxes having a length limit of three characters. This did not impact our analyses, as participants' estimates were clamped after the experiment so that no estimates were below 0° F or higher than 120° F.

[3] In professional forecasting practice, the predictor values of similar cases could be hand-tuned. For their convenience, our online participants were allowed to choose similar cases (i.e., cities) from a predetermined list.

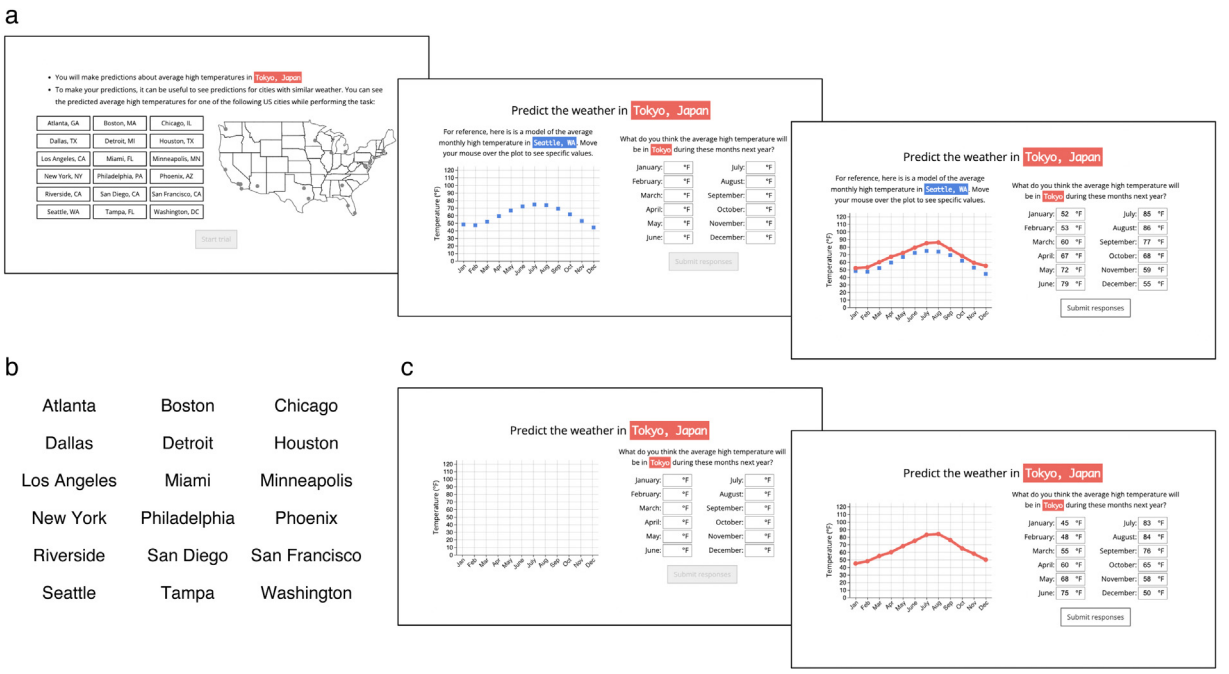


**Fig. 2.** Experimental details. The US-based participants made judgments on three cities: one US city (Kind task environment), one non-US city in the Northern hemisphere (Challenging task environment), and one non-US city in the Southern hemisphere (Wicked task environment). For each city, participants predicted the average high temperature for each month for the target city during the following year. Participants made their estimates for each month by entering numbers in the corresponding text boxes. As participants entered their estimates, a temperature curve of their predictions was drawn on a graph on the left side of the screen. (a) Participants in the Model-Assist condition selected a model city before each trial. After choosing a model city, a polynomial regression model of the selected model city's true average high temperatures was displayed on the same temperature graph that participants' estimates were drawn on. The model city predictions were displayed as a series of points and in a complementary color to the generated graph of participants' estimates. Participants could observe the exact predicted values for each month for the model city by hovering their mouse over the temperature plot. (b) Model cities consisted of the 18 largest US metropolitan areas. (c) Participants in the Control condition completed the same forecasting task as Model-Assist participants, but completed the task on their own without any model assistance.

of the 18 largest US metropolitan areas and are listed in Fig. 2. Upon selecting a model city, Model-Assist participants were shown a fourth-degree polynomial regression model of the model city's true average high temperatures (see Fig. 2 for an example).

All participants first completed a practice trial in which they forecasted high temperatures in Berlin, Germany (Berlin was not a target city and data from this trial were excluded from analysis). Participants in both conditions made these predictions without access to a model city. Following the criteria in our preregistration, participants who made nonsensical temperature predictions on Berlin were excluded, and another participant was recruited in their place until we reached our target sample size (see Appendix A.1). Altogether, we recruited 1636 participants, of whom 196 (11.98%) were excluded, for a total sample size of 1440 participants. After making their judgments on three target cities, participants then completed a 10-question multiple-choice quiz on weather-related topics.

### 3.3. Results

***Task environment manipulation check.*** Our first set of analyses investigated whether prediction errors between participants' estimates and the true average highs for each city differed by task environment as expected (i.e., higher error on Wicked vs. Challenging tasks and on Challenging vs. Kind tasks). We did so by fitting a linear mixed-effects regression model to all participants' judgments with the absolute error (the absolute difference between the predicted high temperature and the true average high temperature for each month/city) as the dependent variable, and the task environment (a factor with levels for Kind, Challenging, or Wicked) as an independent variable. We also included random intercepts for each participant. This model was fit to judgments made by participants in both conditions.

We performed analyses on this model using the `emmeans` package in R to investigate the pairwise differences in the relevant levels of the task environment variable (we use this same approach on the relevant fixed effects for all subsequent analyses). As expected, we found that the average absolute error was significantly higher on cities in the Wicked task environment compared to cities in the Challenging task environment (mean Wicked task environment absolute error: 16.9°; Challenging task environment: 14.1°; $z = 25.39$, $p < 0.001$). Similarly, we found that the average error was significantly higher in cities in the Challenging task environment compared to cities in the Kind task environment (mean Kind task environment city absolute error: 12.1°; $z = 19.23$, $p < 0.001$).

**Table 1**
We categorized target cities by how challenging they are expected to be for the performance of model-assisted judgmental bootstrapping in the kind/wicked nomenclature for task environments (Hogarth et al., 2015). In the Kind task environment, a US model assists with predictions about US cities. In the Challenging task environment a US model assists with predictions about non-US cities in the Northern hemisphere. In the Wicked task environment a US model assists with predictions about cities in the Southern hemisphere that are seasonally reversed from the US model predictions. Each participant made 12 judgments (one for each month) for one city in each of the three task environments.

| Kind task environment<br>*Same country* | Challenging task environment<br>*Different country, same hemisphere* | Wicked task environment<br>*Different country, seasonal reversal* |
|---|---|---|
| Baltimore | Cairo | Auckland |
| Charlotte | Delhi | Buenos Aires |
| Denver | Lagos | Johannesburg |
| Orlando | London | Lima |
| Portland | Mexico City | Luanda |
| Sacramento | Paris | Santiago |
| San Antonio | Tokyo | São Paulo |
| St. Louis | Toronto | Sydney |

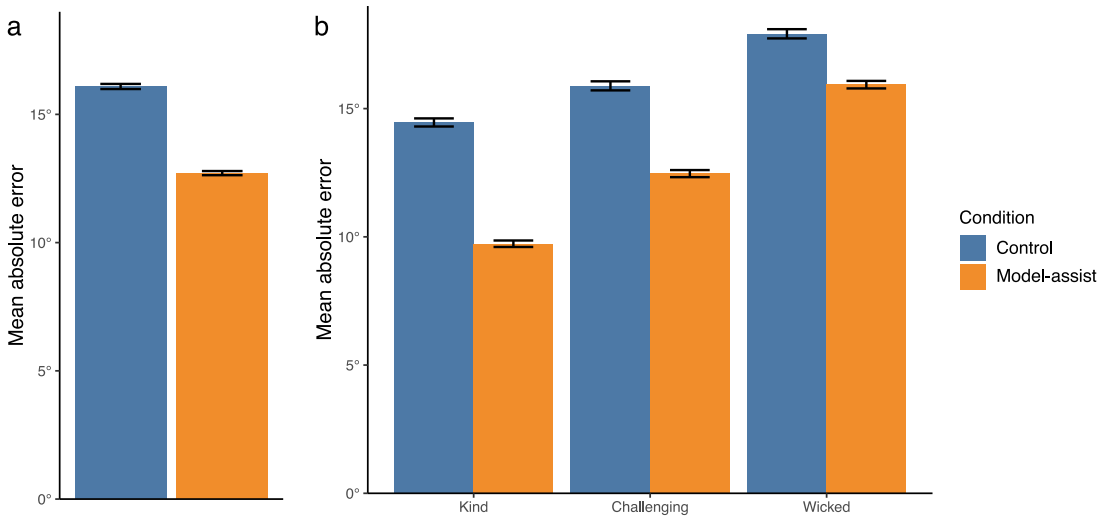


**Fig. 3.** Experiment results. (a) Average absolute error by condition, averaged over participants, task environments, and cities. (b) Average absolute error by condition and domain, averaged over participants and cities. Error bars show standard errors of the means.

***Impact of model assistance.*** Our second set of analyses explored the impact of model assistance on participants' prediction errors. To investigate, we fit a linear mixed-effects regression model with the absolute error as a dependent variable, and fixed effects for the condition (Control and Model-Assist), task environment (Kind, Challenging, or Wicked), and all interactions. This model also included random intercepts for each participant, target city, and their interaction. We constructed this model based on the assumption that participants have different abilities, that certain cities are harder than others, and that participants' knowledge varies from city to city.

We first performed planned contrasts comparing error rates in the Model-Assist and Control conditions. As expected, participants in the Model-Assist condition made judgments that were, on average, significantly more accurate than those in the Control condition: Model-Assist participants' judgments had an average absolute error of 12.7°, compared to 16.1° for participants in the Control condition ($z = 6.03$, $p < 0.001$; see Fig. 3). Furthermore, we observed this same pattern for every target city, with a lower average error among Model-Assist participants compared to participants in the Control condition (see Fig. A.1).

We then used our mixed-effects regression model to investigate the effects of model assistance by task environment. As predicted, we found that model assistance reduced average error for the Kind (Model-Assist error: 9.7°, Control error: 14.5°; $z = 5.22$, $p < 0.001$) and Challenging (Model-Assist error: 12.5°, Control error: 15.9°; $z = 4.69$, $p < 0.001$) task environments, and a post hoc test found the same for the Wicked tasks (Model-Assist error: 15.93°, Control error: 17.92°; $z = 3.29$, $p = 0.001$). However, the benefits of model assistance decreased for more difficult task environments: the mean absolute prediction error for Model-Assist participants was 4.77° lower than Control participants on cities in the Kind task environment, 3.4° lower on Challenging task environment cities, and just 2° lower on Wicked task environment cities. Preliminary results also suggest that model assistance is most useful to less-knowledgeable

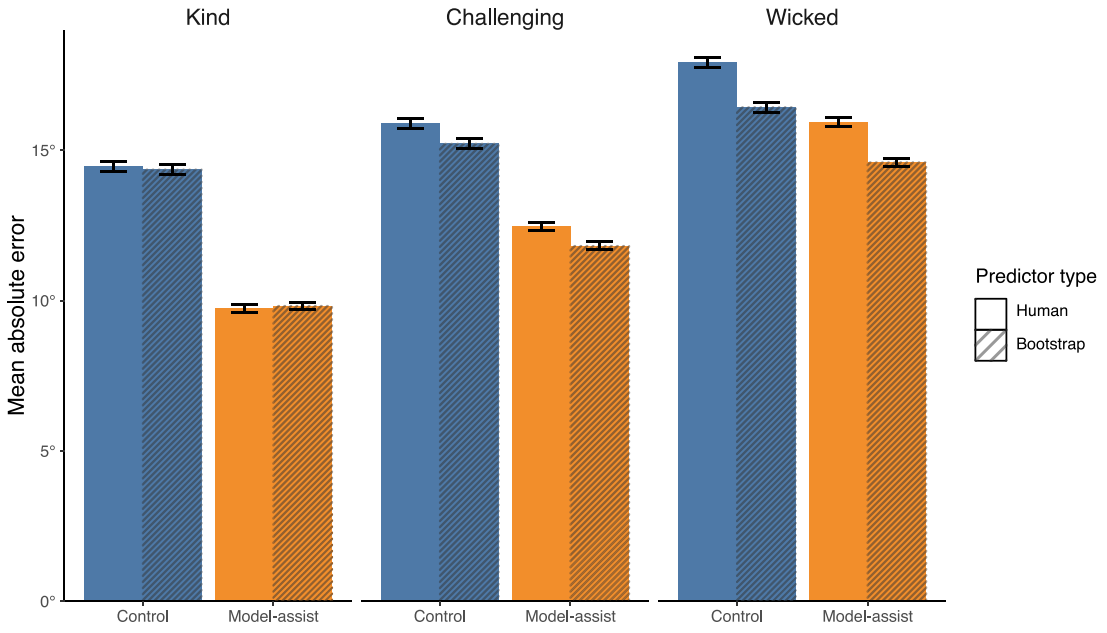


**Fig. 4.** Bootstrapping results by condition and domain. Solid bars show the average absolute error of participants' estimates, and striped bars the average absolute error of bootstrapped models fit to the estimates of the participants in the relevant condition and domain. Error bars show standard errors of the means.

participants, and its effects decrease as participants' abilities increase (see the Discussion and Fig. A.2). However, even for expert participants in the Wicked task environment, model assistance never reduced participants' accuracy, and at worst had no effect.

***Impact of judgmental bootstrapping.*** Finally, we investigated the effects of judgmental bootstrapping on prediction errors to examine the effect of modeling participants' new-domain estimates as compared to using the raw new-domain predictions that they provided. To do so, we fit a separate fourth-degree ridge ($L^2$-norm) polynomial regression model to each participant's temperature predictions for each trial (i.e., each set of 12 monthly temperature predictions on a target city; in total, we fit $1440 \times 3 = 4320$ separate bootstrapping models). We chose a fourth-degree polynomial regression model to strike a balance between modeling the nonlinear seasonal patterns in temperature data and preventing overfitting. Seasonal temperatures are generally well captured by second- or third-degree polynomials, and using a fourth-degree polynomial ridge regression allowed us to capture these patterns in a common model. For consistency with the constraints placed on participants, all bootstrap predictions were rounded to the nearest whole degree.[4]

To select the penalty parameter $\lambda$ for each ridge regression model, we ran leave-one-out cross-validation (LOO CV) on the data from each trial (i.e., the set of 12 temperature estimates a participant made for a given target city). We then followed the one standard error rule (Chen & Yang, 2021) and selected the largest $\lambda$ with a CV error within one standard error of the lowest CV error. To facilitate numerical stability with minimal changes to the predicted values, we added a small amount of random noise ($\epsilon \sim \mathcal{N}(0, 0.01^2)$) to participants' estimates on trials with fewer than three unique temperature estimates.

We investigated the effects of bootstrapping separately for participants in both the Model-Assist and Control conditions. For our analysis, we fit a linear mixed-effects regression model predicting the difference in absolute error between participants' predictions and the predictions of the bootstrapped models for each city and month. This model also included the condition as a fixed effect and random intercepts for each trial (i.e., each fitted bootstrapped model).

In line with past work, we found that judgmental bootstrapping significantly reduced the average error in both conditions. On average, bootstrapping reduced the error of both Control participants' estimates (average reduction: 0.75°; $z = 18.59$, $p < 0.001$) and Model-Assist participants (average reduction: 0.63°; $z = 15.80$, $p < 0.001$). As shown in Fig. 4, the direction of this effect complemented the effects of model assistance. Whereas the benefit of model assistance decreased with the difficulty of the task environment, bootstrapping had the largest observed effects on cities in the Wicked task environment, smaller but still positive effects in the Challenging task environment, and no clear effect on Kind task environment cities. In the Appendix, we report results showing that a much simpler bootstrapping model that averages participants' estimates for the preceding and trailing months has similar effects as the polynomial ridge regression model (see Fig. A.3). This suggests that the benefits of bootstrapping can be robust across different models and architectures.

[4] As we did for human judgments, any bootstrap predictions that were greater than 120° were set to 120°, and judgments that were less than 0° were set to 0°.

## 4. Discussion

While model-assisted estimation was found to improve accuracy overall (Fig. 3), it is not *a priori* clear that all parts of the process did so. In this section, we present an exploratory analysis of the two model-assisted estimation steps that involve human judgment (steps a and c in Fig. 1).

### 4.1. How well do people choose old-domain cases?

In the first step of model-assisted estimation, people construct or choose cases to present to the old-domain model to obtain predictions that they will later consult. Fig. 5 shows the most common city that was passed to the old-domain model for each target city, as well as the model city that would have required the least adjustment in terms of the mean absolute error (MAE). For five out of eight target cities in the Kind task environment (the domain requiring the least adjustment), participants' most commonly chosen model city matched the city with the lowest MAE. The differences between the model and target cities were greater in the Challenging domain (especially in Cairo, Delhi, and Mexico City), and participants' most-chosen model city matched the city with the lowest MAE for two of the eight target cities. In the Wicked task environment, the city with the lowest MAE was the model choice for one target city. We note, however, that participants who are knowledgeable about seasonal reversal might intentionally select model cities with a high MAE depending on the transformation they plan to apply. For example, Miami could be an appropriate model city for Sydney if one intends to transform the Miami predictions by flipping them around the January 1st values. Future work should thus develop more sophisticated methods for evaluating participants' chosen models and investigate the effects of these selections on accuracy. Similarly, future work should investigate the heuristics and processes individuals use to select old-domain cases. As suggested above, participants' selections likely depend on similarity estimates along relevant dimensions, their knowledge of both the target and old-domain cases, and the complexity of the transformations they plan on making.

In related work, (Lawrence, Goodwin, & Fildes, 2002) also gave participants a choice of what kinds of model predictions they would like to see. That is, participants in their study chose the forecasting method (e.g., exponential smoothing, Holt's method, etc.) displayed on the screen. They found that giving participants this choice led to less accurate predictions than simply presenting an optimal forecast as determined by the system. By contrast, in our setting, we assume that an optimal forecast cannot be generated because there is no model or outcome data available in the new domain, and the predictor values required to call the old-domain model may not fully align with those in the new domain.

### 4.2. How well do people make judgmental adjustments?

The final step of the model-assisted estimation process is judgmental adjustment, in which experts adjust the outputs of the old-domain model for the new domain (step c in Fig. 1). Prior research has found that such adjustments are often not helpful. In the words of Lim and O'Connor 1995 the "finding that people could benefit from the reliable [i.e., accurate] model, but did not outperform it, has commonly been shown". The left panel of Fig. 6 shows the results of an exploratory analysis comparing the error of participants' adjusted predictions to the error that would result if no adjustments were made to the predictions of the old-domain model. Participants' adjustments in the Kind task environment reduced accuracy. In the Wicked task environment, adjustment had little effect, and the impact in the Challenging task environment was between these two outcomes. Participants benefited from more accurate predictions (that is, they had relatively lower errors in the Kind task environment in which old-domain model predictions were more accurate) but were not able to improve upon them. Consistent with prior findings, participants' adjustments to accurate predictions only made things worse (Lawrence et al., 2006; Lim & O'Connor, 1995; Willemain, 1989)

Presumably, individuals with greater expertise in climate science and geography would be able to make beneficial adjustments, most obviously by accounting for seasonal reversal in the Wicked task environment. We repeated the exploratory analysis above on participants with abilities in the top 10% of climate knowledge, as per an item-response theory (IRT) model fit to the weather knowledge quiz to estimate participant expertise (see Appendix A.3). In the Wicked task environment, these experts' adjustments greatly reduced the MAE compared to the cities they chose. In the Challenging task environment, the expert adjustments had little effect, and in the Kind task environment, experts' adjustments slightly reduced the accuracy, consistent with prior work (e.g., Lim & O'Connor, 1995).

Because participants decide on the cases they would like to submit to the old-domain model, they might have good intuitions about when the old-domain predictions will be particularly appropriate for the target case in the new domain. This would often be the case in a Kind task environment: one might know that Baltimore's predicted weather could be a stand-in for that of Washington DC. A future direction for improving model assistance could be to make it easy for experts to submit unadjusted old-model predictions, or to downweight their small adjustments. Alternatively, experts could simply be discouraged from making small adjustments, a practice that, according to Fildes and Goodwin (2007), would save time and not harm accuracy. Future work should test these methods and investigate the processes participants follow when making adjustments.

While we found that judgmental adjustments sometimes reduce predictive accuracy, we note that the same cannot be said of model-assisted estimation in general, which consistently proved beneficial (see Fig. 4 and Fig. A.1). This is because judgmental adjustment is just one part


**Fig. 5.** Selected cities for Model-Assist participants by target city. Each plot shows (1) the average highs of the target city in black, (2) the average highs of the most commonly chosen model city in teal, and (3) the average highs for the model city with the lowest mean absolute error in red. All temperature curves are plotted by month, from January to December. The top row shows cities in the Kind task environment, the middle the Challenging task environment, and the bottom the Wicked task environment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

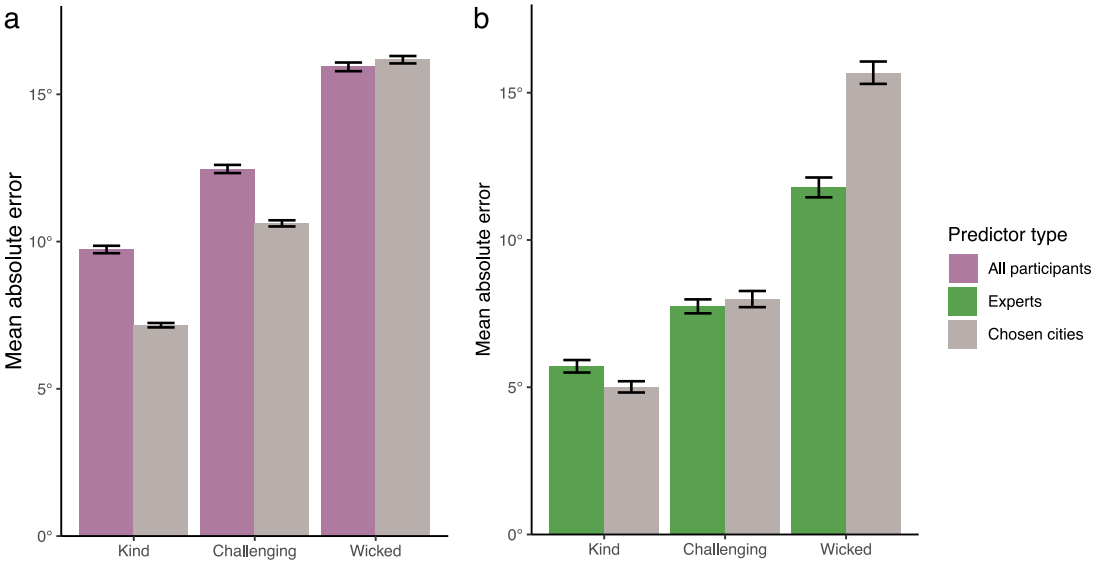


**Fig. 6.** Comparison of Model-Assist participants and their chosen model cities. (a) Average absolute error of Model-Assist participants' predictions compared with the average absolute error of the chosen cities. The average error of the model city was computed using the polynomial regression model of the city's true high temperatures that was shown to the participant as the predicted values for the target city. (b) Average absolute error of expert Model-Assist participants compared with the average absolute error of the models of the experts' chosen cities. We defined experts as those with abilities in top 10% based on our item-response theory (IRT) model (see Appendix A.3). Error bars show standard errors of the means.

of model-assisted estimation. In general, the benefits of providing experts with an old-domain model and allowing them discretion to select cases significantly outweighs the harms sometimes caused by judgmental adjustment.

## 5. Conclusion

In this work, we proposed and tested a method called model-assisted judgmental bootstrapping for training models to make out-of-population predictions in new domains that lack outcome data. Model-assisted judgmental bootstrapping consists of two main steps: (i) model-assisted estimation, where experts construct or choose cases for which they would like to see predictions from an old-domain before making estimates for a new domain; and (ii) judgmental bootstrapping, where statistical models are fit to experts' judgments and then used to make predictions for new cases. This combined approach leverages the strengths of human judgment as well as statistical modeling.

We conducted a large preregistered experiment to assess how the components of the method affect predictive accuracy across increasingly difficult task environments. As predicted, both model assistance and judgmental bootstrapping significantly improved accuracy. Importantly, their effects were complementary. Model assistance had the greatest impact in the Kind task environment (where the new domain was similar to the old domain), and the least in the Wicked task environment (where the new domain was most different compared to the old domain). For judgmental bootstrapping, this pattern was reversed: it provided the largest benefits in the most difficult task environment. Together, these effects combined so that the method led to similar reductions in absolute error in each task environment; on average, model-assisted judgmental bootstrapping reduced the error by 4.68° in the Kind task environment, 4.1° in the Challenging task environment, and 3.3° in the Wicked task environment (see Fig. 4).

Regarding limitations, our study was exclusively focused on the domain of temperature forecasting, a specific type of time series prediction, but left unexplored many other kinds of forecasting tasks. Additionally, the distinction between novices and experts in our research was simulated by varying task environments for participants who generally lacked experience in predictive modeling. Field tests with actual experts could provide a more accurate assessment of the method's practical value. Our method as stated enables experts to engage creatively with the existing domain model to generate useful predictions, such as by manually specifying and changing predictor values. However, our empirical test was greatly simplified, allowing participants only to select a model city. To test how well the model performs in realistic conditions, a future study should involve expert participants who are empowered to choose old domains and call models in any manner they deem appropriate. Another worthwhile test would involve exploring new domains where trained models and outcome data are entirely absent. In our study, models and outcome were only effectively absent in the sense that participants did not have direct access to them. Similarly, the empirical tests could also be made more challenging by using less accurate old-domain models. Finally, the scope of our study was specifically focused on a scenario in which the old and new domain outcome variables were very similar (average high temperatures). Given that our method can accommodate scenarios in which outcomes in the old and new domains are related yet distinct—for example, applying a model of unemployment rates in one domain to assist in estimating demand for social assistance in another—it will be important in future research to evaluate its effectiveness in such scenarios.

In many forecasting tasks, model assistance should be relatively simple to incorporate; experts only need to identify and consult an existing model trained on a different domain. Leveraging judgmental bootstrapping may be significantly more difficult, as forecasters must choose the architecture of the bootstrapping model, make several predictions for the new domain, and train the final model. However, preliminary results from our experiment suggest that even very simple bootstrapping models can be effective. As shown in Fig. A.3, we found that a model that simply averages participants' estimates for the preceding and trailing months had similar beneficial impacts as fitting a polynomial ridge regression model. This finding connects with research finding that simple weighting schemes performed similarly to multiple regression models (Dawes & Corrigan, 1974), that simple combinations of judgments performed as well as more sophisticated ones (Edmundson, 1990; Lawrence, Edmundson, & O'Connor, 1986), and that averaging an individual's predictions for the same problem can improve predictive accuracy (Herzog & Hertwig, 2014). Forecasters may thus find that even highly simple bootstrapping models that are fit on only a handful of estimates may be valuable tools for improving accuracy.

Of all the results of this investigation, we find the complementary contribution of model assistance and judgmental bootstrapping particularly intriguing, as this pattern stabilizes the method's performance across a variety of environments. If this compensatory pattern replicates in future work, our method could be well suited for scenarios in which forecasters do not have a good sense for how difficult the task environment is.

## CRediT authorship contribution statement

**Mathew D. Hardy:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Sam Zhang:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Jessica Hullman:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Jake M. Hofman:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Conceptualization, Methodology, Formal analysis. **Daniel G. Goldstein:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data and code availability

Data and code from our experiments are available in a public GitHub repository at https://github.com/mdahardy/judgmental-bootstrapping.

## Appendix

### A.1. Participant filtering

As specified in our preregistration, we filtered participants based on their responses to an initial practice trial to exclude bots and highly inattentive participants. During this practice trial, participants were asked to forecast the high temperatures in Berlin. All participants completed this task without model assistance.

We used the following criteria to ensure each participant has at least a basic level understanding of climate and temperature forecasting:

- **Seasonal consistency:** We required each participant's forecasted temperature for January to be lower than the forecast for July. This ensured that participants have a basic understanding of seasonal temperature variations and the difference between winter and summer.
- **Bounded but variable estimates:** We required the distance between a participant's highest and lowest estimates to be at least 2° and no greater than 100°. Additionally, all of a participant's predictions were required to fall between 0° and 120° Fahrenheit.
- **Reasonable accuracy:** We required that the mean absolute difference between each participant's estimates and the true average high temperatures should be less than or equal to 50° Fahrenheit.

To reduce automated bot submissions, we also incorporated "honeypot" bot instructions during the Berlin practice trial. These instructions were discernible only to agents directly interacting with the experiment programmatically, rather than the rendered webpage. These instructions told participants that were paying attention to respond with a temperature of 99° for every month (bots following these instructions would be excluded using the criteria outlined above).

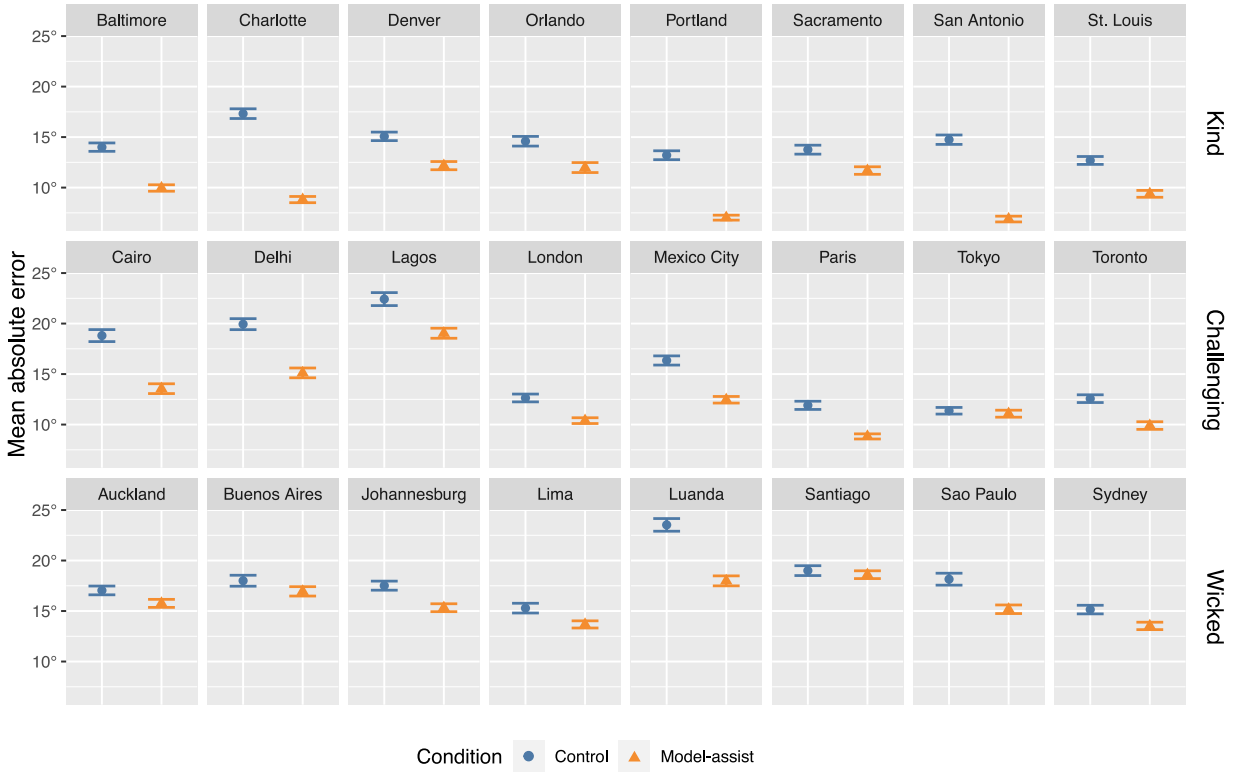


**Fig. A.1.** Effects of model assistance by target city. Each plot shows the average absolute error for participants' predictions in both conditions (Control and Model-Assist) on that city. Error bars show standard errors of the means. The top row shows cities in the Kind task environment, the middle the Challenging task environment, and the bottom the Wicked task environment.

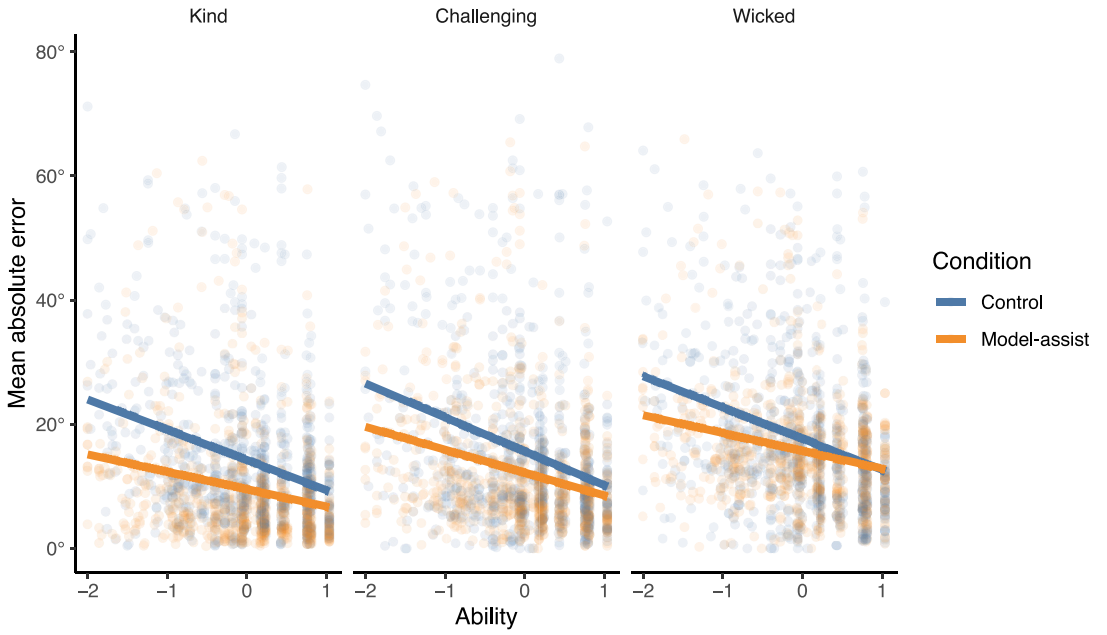


**Fig. A.2.** Average error by condition and estimated ability. Each point shows the participant's estimated ability and the average absolute error of their predictions on the city for the relevant domain. Participant ability is taken from an item-response theory (IRT) model fit to the participant's responses to a post-experiment comprehension quiz. Each participant is plotted once on each panel (Kind, Challenging, and Wicked task environments), corresponding to the three domains they made predictions on. Points are colored based on the participant's condition. Lines show fitted linear regression lines for the relevant data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

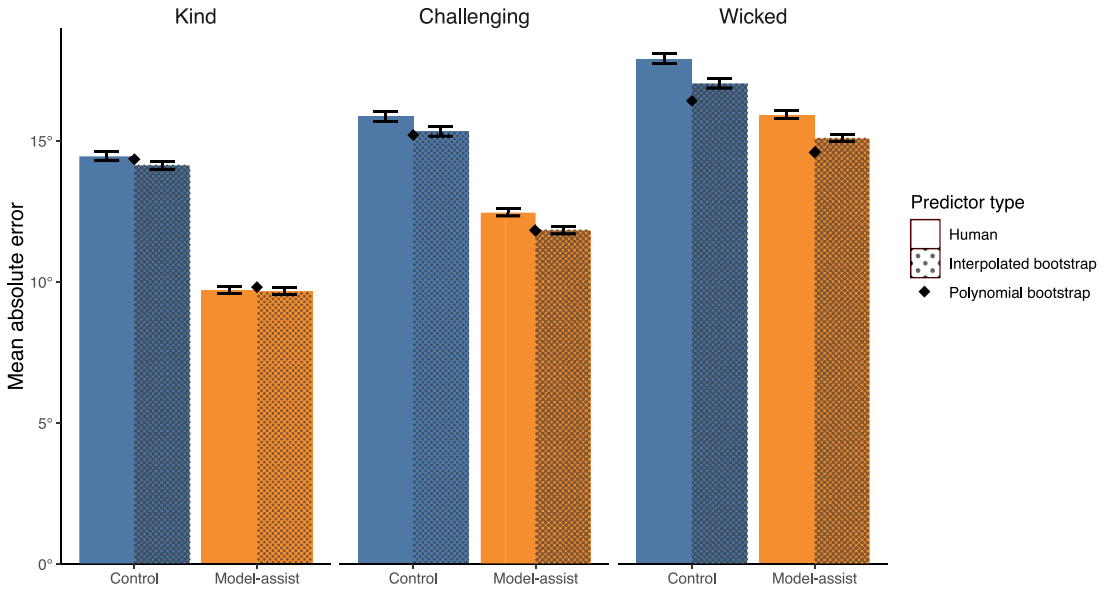


**Fig. A.3.** Effects of simple interpolation by condition and domain. Solid bars show the average absolute error of participants' estimates, and dotted bars the average absolute error of interpolated estimates for the relevant condition and domain. Interpolated estimates for participants' estimates were fit by averaging the participant's predictions for the preceding and trailing month (with December preceding January) for the same city. Diamond dots show the average bootstrap error for each domain and condition (plotted in Fig. 4), and error bars show standard errors of the means.

Participants who failed to meet these criteria were excluded from our analyses, and we recruited a new participant in their place until we reached our target sample size. While excluded from our analyses, filtered participants were not excluded from the experiment and received full payment for the HIT.

### *A.2. Power analysis*

We determined our sample size using a power analysis on simulated data of our proposed experiment. These simulations involved repeatedly generating experiment

datasets and then running our preregistered analyses on the generated data. We describe this process below.

We first generated simulated datasets by sampling with replacement from pilot data of our experiment. To do so, we first sampled equal numbers of participants with replacement from both conditions (720 participants for both Control and Model-Assist). As participants in pilots made judgments on more than three cities, we then randomly sampled a single city for each task environment (Kind, Challenging, and Wicked) for each participant. This setup ensured that our generated dataset matched the structure and size of our proposed experimental design.

After constructing a dataset, we performed each of our preregistered analyses on the generated data. We independently repeated this process 100 times. In 80 of these 100 simulations, all of our hypotheses were as expected, leading us to determine that our proposed design was sufficiently powerful.

### A.3. IRT model

After making their judgments, each participant completed a 10-question multiple-choice quiz on weather-related topics before finishing the quiz (questions are included in our GitHub repo). After the experiment, we fit a three-parameter item-response theory (IRT) model to participants' quiz choices to determine the "ability" of each participant using the `ltm` package in R. These ability estimates were approximately normally distributed, with higher values indicating higher ability (ability mean: −0.06; standard deviation: 0.758; min: −2.00; max: 1.04).

## References


Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*: *vol. 30*, Springer.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, *27*(3), 411–422.

Chen, Y., & Yang, Y. (2021). The one standard error rule for model selection: Does it work? *Stats*, *4*(4), 868–892.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106.

Di Cecco, G. J., & Gouhier, T. C. (2018). Increased spatial and temporal autocorrelation of temperature under climate change. *Scientific Reports*, *8*(1), 14850.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Edmundson, R. (1990). Decomposition; a strategy for judgemental forecasting. *Journal of Forecasting*, *9*(4), 305–314.

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial college COVID-19 response team. *Imperial College COVID-19 Response Team*, *20*(10.25561), 77482.

Fildes, R., & Goodwin, P. (2007). Good and bad judgement in forecasting: Lessons from four companies. *Foresight*, *8*(Fall), 5–10.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, *12*(1), 37–53.

Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, *9*(2), 147–161.

Harvey, N. (1988). Judgmental forecasting of univariate time series. *Journal of Behavioral Decision Making*, *1*(2), 95–110.

Harvey, N., Harries, C., & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes*, *81*(2), 252–273.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, *24*(5), 379–385.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies.. *Psychological Bulletin*, *134*(3), 404.

Katsikopoulos, K. V., Simsek, O., Buckmann, M., & Gigerenzer, G. (2021). *Classification in the wild: The science and art of transparent decision making*. MIT Press.

Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1986). The accuracy of combining judgemental and statistical forecasts. *Management Science*, *32*(12), 1521–1532.

Lawrence, M., Goodwin, P., & Fildes, R. (2002). Influence of user participation on DSS use and decision accuracy. *Omega*, *30*(5), 381–392.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518.

Lim, J. S., & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making*, *8*(3), 149–168.

Remuzzi, A., & Remuzzi, G. (2020). COVID-19 and Italy: what next? *The Lancet*, *395*(10231), 1225–1228.

Sanders, N. R., & Ritzman, L. P. (2001). Judgmental adjustment of statistical forecasts. In *Principles of forecasting: A handbook for researchers and practitioners* (pp. 405–416). Springer.

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. OUP USA.

Willemain, T. R. (1989). Graphical adjustment of statistical forecasts. *International Journal of Forecasting*, *5*(2), 179–185.

Yntema, D. B., & Torgerson, W. S. (1961). Man-computer cooperation in decisions requiring common sense. *IRE Transactions on Human Factors in Electronics*, (1), 20–26.