ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Decomposing the effects of crowd-wisdom aggregators: The bias–information–noise (BIN) model

Ville A. Satopää [a,*], Marat Salikhov [b,c], Philip E. Tetlock [d], Barbara Mellers [d]

[a] INSEAD, France
[b] Moscow School of Management Skolkovo, 100 Novaya Street, Skolkovo, Moscow Region, 143026, Russia
[c] New Economic School, 3 Nobelya Street, Moscow, 121205, Russia
[d] The Wharton School of the University of Pennsylvania, Philadelphia, PA, United States of America

ARTICLE INFO

Keywords:
Judgmental forecasting
Partial information
Prediction markets
Wisdom of crowds
Bayesian Statistics Shapley Value

ABSTRACT

Aggregating predictions from multiple judges often yields more accurate predictions than relying on a single judge, which is known as the wisdom-of-the-crowd effect. However, a wide range of aggregation methods are available, which range from one-size-fits-all techniques, such as simple averaging, prediction markets, and Bayesian aggregators, to customized (supervised) techniques that require past performance data, such as weighted averaging. In this study, we applied a wide range of aggregation methods to subjective probability estimates from geopolitical forecasting tournaments. We used the bias–information–noise (BIN) model to disentangle three mechanisms that allow aggregators to improve the accuracy of predictions: reducing bias and noise, and extracting valid information across forecasters. Simple averaging operates almost entirely by reducing noise, whereas more complex techniques such as prediction markets and Bayesian aggregators exploit all three pathways to allow better signal extraction as well as greater noise and bias reduction. Finally, we explored the utility of a BIN approach for the modular construction of aggregators.

1. Introduction

Many studies have demonstrated the power of aggregating predictions to improve their accuracy (for excellent reviews, see Armstrong 2001, Clemen 1989, Clemen and Winkler 1999, Winkler, Grushka-Cockayne, Lichtendahl Jr, and Jose 2019). These studies explored a wide range of methods for aggregating predictions, such as machine learning and statistical techniques (Makridakis, Spiliotis, & Assimakopoulos, 2018; McAndrew, Wattanachit, Gibson, & Reich, 2021), and prediction markets (PMs) (Chen & Pennock, 2010; Elliott & Timmermann, 2013). Previous studies also examined methods for selecting and assessing forecasters (Budescu & Chen, 2015; Mannes, Soll, & Larrick, 2014), organizing interactions (Tetlock & Gardner, 2016), and sharing information (Palley & Soll, 2019; Satopää, Pemantle, & Ungar, 2016; Sunstein & Hastie, 2015). In numerous fields, interest is growing in moving beyond horse race comparisons of the aggregator that can best reduce the error scores to identifying the mechanisms that allow promising methods to improve the accuracy of predictions (Davis-Stober, Budescu, Dana, & Broomell, 2014; Erev, Wallsten, & Budescu, 1994; Soll & Larrick, 2009).

Recently, Satopää, Salikhov, Tetlock, and Mellers (2021) proposed a new statistical framework called the bias–information–noise (BIN) model for explaining how experimental treatments in forecasting tournaments can improve the accuracy of individual forecasters via three pathways: reducing random error or noise, decreasing systematic error or bias, and boosting signal/information extraction. The authors used the BIN model to dissect the

* Corresponding author.
E-mail address: ville.satopaa@insead.edu (V.A. Satopää).

https://doi.org/10.1016/j.ijforecast.2021.12.010
0169-2070/

effects of tournament interventions, such as training in probabilistic reasoning, working in teams, and tracking talented forecasters and then placing them in elite teams.

However, Satopää et al. (2021) focused solely on modeling the treatment effects on individual forecasters, rather than the strengths and weaknesses of aggregators. By contrast, in the present study, we focused exclusively on aggregators. By analyzing aggregation techniques in the same manner as (Satopää et al., 2021) analyzed experimental treatments, we aimed to move beyond tabulating the outcomes of horse races to obtain a deeper understanding of aggregators. A more apt metaphor is a professional car race where mechanics can look under the hood and inspect the performance engines that underlie aggregators. We applied a range of aggregators to a series of probability predictions in forecasting environments with variations in terms of the size of the crowd, level of skill and information sharing within the crowd, and forecast time horizon. We aimed to understand the pathways that allow different types of aggregators to synthesize wisdom from crowds.

This granularity is important for the following two reasons. First, predictions made in different environments are likely to have different levels of bias, noise, and information asymmetry. Thus, our choice of aggregation tool should match the environment. For instance, a potentially steep accuracy cost may be incurred if we apply a pure noise reduction aggregator to predictions that vary due to information asymmetry (different forecasters possess distinct sets of valid knowledge). Second, we can use BIN decomposition to advance the modular construction of aggregators, which involves partitioning an aggregator into modules, inspecting how each module contributes to the overall accuracy, and detecting redundancies and potential for improvements.

2. Bias, information, and noise

The BIN model developed by Satopää et al. (2021) is an analytic engine that traces the variations in observable forecasting accuracy to three unobservable, mutually exclusive and exhaustive components: systematic error (bias), random error (noise), and lack of knowledge (partial information). The model locates these components in the framework of a signal universe that contains all past and future signals with positive or negative effects on a target event. The model treats these signals as causal, where the target event occurs if and only if the cumulative signal contribution is positive. In addition to relevant signals, the universe contains signals that are not relevant to the event. Forecasters sample and interpret all of these signals with various levels of skill and thoroughness. Their samples consist of relevant signals (partial information) and irrelevant signals (noise). The BIN model makes the bounded rationality assumption that forecasters have a limited ability to distinguish relevant from irrelevant signals in their samples. Irrelevant signals increase noise and relevant signals increase information. In addition, forecasters may fail to center the signals incorrectly, thereby leading to systematic biases. For a stylized example, see Example 1 given by Satopää et al. (2021).

Each pathway has a distinctive observable effect on the predictions made by forecasters. As an illustrative example, we used the BIN model to simulate outcomes with a base rate of 0.5 and forecasters' predictions with different levels of bias, noise, and information. Fig. 1 shows calibration plots that graphically illustrate these predictions against the objective frequency of event occurrence. To track whether these predictions are extreme or confident, the inset boxes in the bottom right corners show histograms representing the predictions made by the forecaster for a large number of events.

Panel A depicts a perfectly calibrated forecaster without bias or noise. The calibrated forecaster's probability predictions align perfectly with the objective frequencies of events. For example, whenever the forecaster reported a probability of 0.2, then exactly 20% of the events occurred, and the equivalent relationship also applied to all possible predictions between 0 and 1 (not only 0.2). Panel B illustrates the marginal effect of increasing information by allowing the forecaster to condition on a variable that correlates more strongly with the outcome. Given that there is still no bias or noise, the predictions still align perfectly with the objective frequencies of events. However, the predictions in panel B are more extreme or decisive, thereby indicating their greater resolution than those in panel A. The panel B forecaster would be more valuable to decision makers because this forecaster has all three advantages of BIN: possession of relevant information and absence of bias and noise.

The remaining four panels illustrate the effects of bias and noise. Panel C depicts a forecaster who makes systematically low predictions and who has a downward bias. Panel D is a forecaster with an upward bias and panels E and F represent noisy forecasters. The difference between panels E and F is that in panel F, the forecaster's irrelevant signals are non-negatively correlated with the forecaster's relevant signals, thereby making the forecasters' predictions more extreme than can be justified. By contrast, in panel E, irrelevant signals are negatively correlated with the forecaster's relevant signals, which leads to predictions that are excessively close to the non-informative base rate of 0.5, and thus they are under-confident. In panels C, D, E, and F, the predictions no longer align with the objective frequencies of events, which is a sign of miscalibration. The panels show how bias shifts the points up or down, and noise rotates the points either clockwise or counterclockwise.

3. Overview of the BIN model

In this section, we briefly review the technical details of the BIN model. We denote the target event as $Y \in \{0, 1\}$ such that $Y = 1$ if the event happens and $Y = 0$ if it does not. The outcome is determined by a hypothetical normally distributed variable Z^* representing the accumulation of all relevant signals in the signal universe. The outcome itself is $Y = \mathbb{1}(Z^* > 0)$, where Z^* has mean $\mathbb{E}[Z^*] = \mu^*$ and the indicator function $\mathbb{1}(E)$ equals 1 if E is true; otherwise, 0.

The BIN model treats individual forecasters as exchangeable, which means that the expected levels of bias,

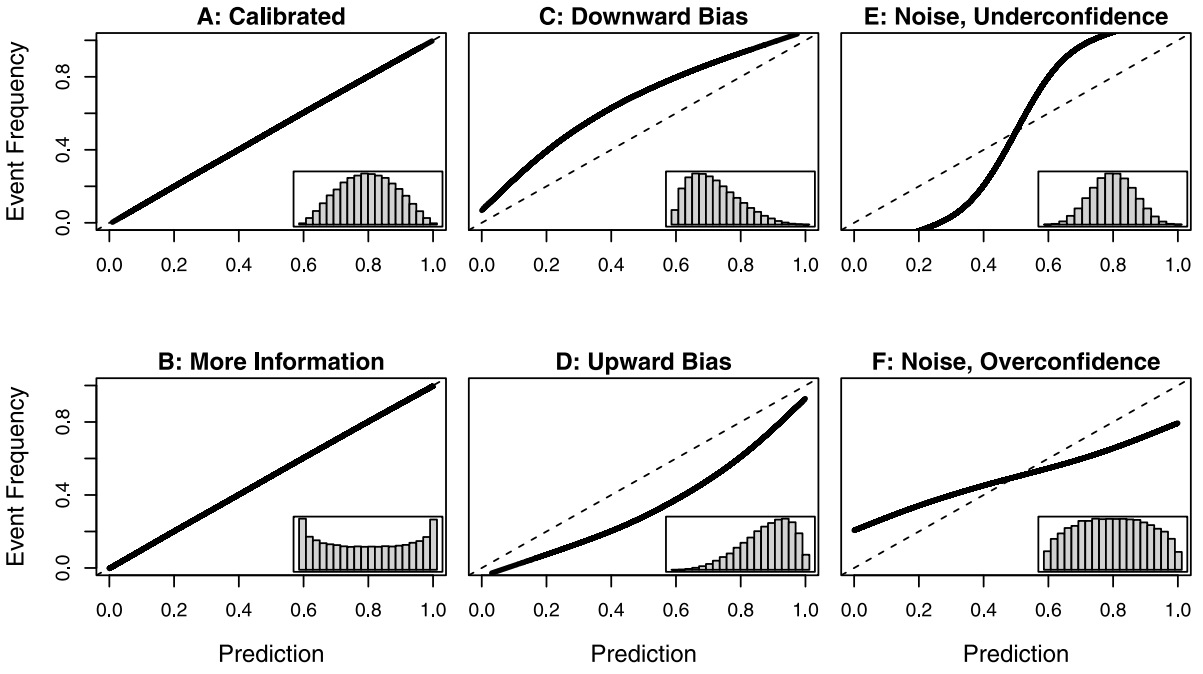

Fig. 1. From left to right, columns show the effects of information, bias, and noise, respectively. The inset figures in the bottom right corners show the distributions of the predictions. Each distribution has its own x-axis defined by its bottom border and it ranges from 0 (left) to 1 (right).

information, and noise are the same for all individuals in the same group. The kth forecaster assigns a probability $p_k \in (0, 1)$ to the event $\{Z^* > 0\}$ based on a normally distributed variable Z_k representing the accumulation of (relevant or not) signals in forecaster k's sample. If the mean of the forecaster's accumulated signals is $\mathbb{E}[Z_k] = \mu^* + \mu$, then the bias is $\mathbb{E}[Z_k] - \mathbb{E}[Z^*] = \mu$. Partial information is the covariance between Z_k and Z^*: $\mathrm{Cov}(Z^*, Z_k) = \gamma$. Noise is the remaining variability of Z_k after removing the covariance with Z^*: $\mathrm{Var}(Z_k) - \mathrm{Cov}(Z^*, Z_k) = \delta$. To map the forecasters' accumulated signals to the probability predictions while ensuring that the model parameters are statistically identifiable, the BIN model makes the bounded rationality assumption that forecasters are not aware of the noise and bias in their accumulated signals Z_k, and they consider that $\delta = 0$ and $\mu = 0$. With the potential for misbelief, the forecaster predicts:

$$p_k = \Phi\left(\frac{Z_k}{\sqrt{1-\gamma}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

In this study, we modeled the aggregator as the individual predictions described above, except that the aggregator was treated as a group of its own.[1] In the same manner as the individual predictions, the aggregate forecast was based on a normally distributed variable representing the accumulation of all signals revealed to the aggregator by the predictions of individual forecasters. This variable was then converted into a probability prediction under the same bounded rationality assumption described above.

The BIN model can be estimated ex-post from predictions and outcomes of multiple events. Under the estimated model, we can compare the expected accuracy of the aggregator against the expected accuracy of individual forecasters and perform marginal analysis to explain the performance differentials in terms of the reduction of noise, decrease in bias, or effective use of dispersed information in crowd predictions. In this manner, we can explore the pathways that allow aggregators to derive wisdom from crowds. In the following sections, we described our investigations based on series of subjective probability predictions made in different forecasting environments.

[1] The aggregate is based on predictions that can become linearly dependent, thereby leading to a degenerate distribution. For example, suppose that we model the predictions of 10 forecasters jointly with their equally weighted average aggregate. Given that the aggregate is an exact linear combination of the predictions, the individual predictions and aggregator span only 10 dimensions rather than 11. As a result, the joint distribution is degenerate, which causes estimation problems. A common solution involves breaking this exact linear relationship by adding a small amount of mean-zero noise to the aggregate prediction. In machine learning, adding noise to the training data often improves convergence and reduces the generalization error (Bishop, 1995). For instance, Cross, Ramos, Mellers, Tetlock, and Scott (2018) "blurred away" the discreteness of the data by adding a small amount of white noise to each prediction. In this study, we added normally distributed noise with mean 0 and standard deviation 0.2 to each individual prediction and aggregator in the probit space, which increased the noise level of the crowd and the aggregator by a small amount. However, this increase occurred for both the crowd and aggregator, so it was canceled out when we compared the aggregator to the crowd and computed the difference between the noise levels. Thus, adding noise allowed the estimation procedure to converge, but it did not noticeably affect the results.

4. Individual forecasters and events

We used the BIN model to analyze aggregators in a 4-year series of probability estimates made by forecasters in geopolitical tournaments sponsored by the research branch of the US intelligence community: Intelligence Advanced Research Projects Activity (IARPA).[2] The full data set includes hundreds of forecasting questions, outcomes, and probabilistic predictions made by the thousands of participants in the Good Judgment Project (GJP). For instance, an illustrative question posed on September 1, 2011 asked whether Serbia would be granted European Union candidacy in the next four months. The question was answered as "no" because the event did not occur by that date.

Provided that a question remained open, forecasters were encouraged to update their predictions in response to new information. If forecasters did not update on a given day, we assumed that their beliefs had not changed. To avoid infinite probit scores in the BIN model, all predictions of exactly 0 or 1 were transformed to 0.01 and 0.99, respectively. Given that the questions were open for varying periods and not all forecasters predicted all events, the numbers of questions and forecasters per question differed across forecast time horizons and experimental conditions. For more details, see section S1 in the Supplementary Material.

Our analysis focused on questions with binary outcomes (yes/no) that remained open for no more than 180 days. Thus, we had a large set of comparable questions. To ensure that forecaster bias had a consistent interpretation, we adjusted questions such that "yes" always referred to a change from the status quo. Thus, the forecasters predicted probabilities of change, and bias was either a systematic over- or under-estimation of change.

The GJP assigned forecasters to the following three experimental conditions (the first two assignments were random and the third was based on performance).

- Individuals with and without training in probabilistic reasoning. In the training condition, forecasters completed a tutorial on probabilistic reasoning based on recommendations from the forecast elicitation literature (O'Hagan, Buck, Daneshkhah, Eiser, Garthwaite, Jenkinson, Oakley, & Rakow, 2006).
- Teams with and without training in probabilistic reasoning. Forecasters worked in teams where they could debate each other's predictions. Each forecaster ultimately made their own predictions.
- Superforecasters. The performance of forecasters was tracked over time. At the end of each tournament year, the top 2% forecasters were designated as "superforecasters" and given an opportunity to work together the following year (Mellers, Stone, Murray, Minster, Rohrbaugh, Bishop, Chen, Baker, Hou, Horowitz, et al., 2015). Again, each forecaster ultimately made their own predictions.

[2] The data can be downloaded at https://dataverse.harvard.edu/dataverse/gjp.

Fig. 2 shows the average Brier scores for individual forecasters in different experimental conditions and over different forecast time horizons defined as the number of days between the prediction and event resolution. The left panel is based on all data, whereas the right panel focuses only on data from year 2. We made this distinction because we could analyze PMs based only on data from year 2, where the individuals with and without training formed appropriate control groups. However, all other aggregators could be analyzed across years. Forecasters in different experimental conditions exhibited significant heterogeneity in terms of skill. Both training and teaming improved the individual accuracy. Among the five conditions, superforecasters were the most accurate by a substantial margin.

We aimed to investigate how and to what extent aggregators could improve the average individual-level Brier scores. Our experimental conditions allowed us to assess the relative added value of various aggregators for untrained and trained individuals, untrained and trained teams, and superforecasters. Given that teams engaged in both informational influence (constructive conversations edging closer to the truth, which ramped up the signal) and normative influence (conformity cascades, which ramped up bias and noise), we expected mixed benefits of teaming for all but the most skillfully run teams (which tended to consist of superforecasters). In the following sections, we avoid presenting redundant results and focus on the most informative contrasts. The full results can be found in sections S3 and S4 of the Supplementary Material.

5. Aggregators

We organized aggregators into two categories: (1) unsupervised one-off aggregators that could be applied to an isolated prediction task without past performance data; and (2) supervised aggregators that could be applied only if historical performance data were available.

Unsupervised aggregators: The unsupervised aggregators examined in this study included the following.

1. Probability average (Stone, 1961). In a theoretical study, Satopää (2021a) explained that measures of central tendency treat disagreements among forecasters as noise. However, if disagreement stems from information asymmetry, these aggregators will mistakenly eliminate private information. Given that averaging is a measure of central tendency, it is unlikely to boost accuracy via the information pathway. Bias also presents a problem because a simple average of probabilities is often excessively close to 0.5 (Baron, Mellers, Tetlock, Stone, & Ungar, 2014). Probabilities are bounded within the unit interval, which means that noise will push large probabilities downward close to 1.0 and small probabilities upward close to 0.0.
2. Trimmed probability average (Jose, Grushka-Cockayne, & Lichtendahl Jr, 2014). The trimmed average removes a user-specified percentage of the lowest and highest probabilities before averaging

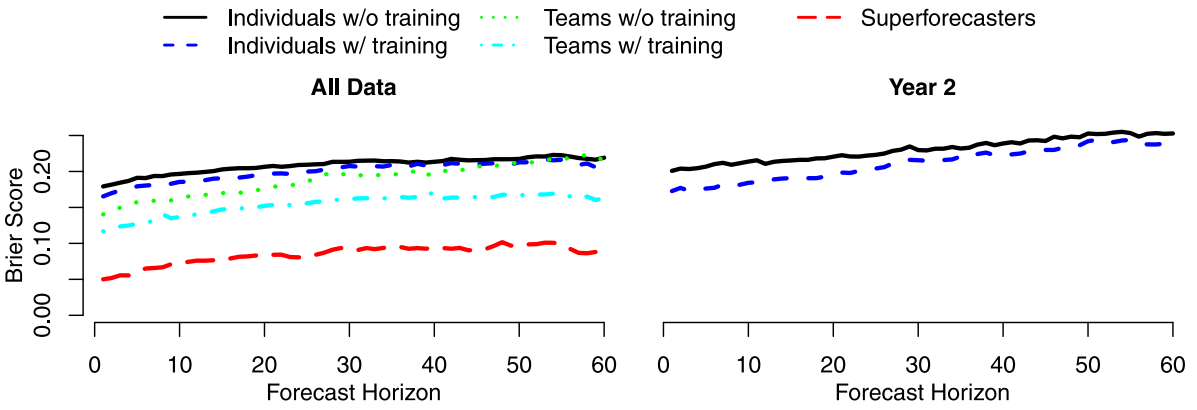

Fig. 2. Average individual level Brier scores in experimental conditions across forecast time horizons.

the remaining probabilities. An alternative to trimming is Winsorizing, which replaces the extreme value with the nearest user-specified percentile instead of removing them. Both approaches can reduce the influence of extreme values and the tendency of the average to be excessively close to 0.5.

3. Median of the probabilities (Hora, Fransen, Hawkins, & Susel, 2013). This aggregator uses extreme trimming because the median excludes all but the middle value of the individual forecasts.
4. Probit average (Satopää, Baron, Foster, Mellers, Tetlock, & Ungar, 2014). To debias the simple probability average, we transformed the probabilities into their probits in the unbounded real line, averaged the probits, and then transformed the average probit back to the probability scale. However, this is still a measure of central tendency. Therefore, similar to the probability average, trimmed probability average, or median, the probit average should be an efficient noise reducer but not an efficient integrator of information dispersed among forecasters.
5. Regularized Bayesian Aggregator (RBA; Satopää 2022). This one-off aggregator can partially separate disagreement due to noise and information asymmetry by exploiting the statistical fact that disagreement among forecasters is limited by the noisiness of their judgments. This method is based on a constrained version of the BIN model. However, unlike the BIN model, the RBA does not incorporate bias or allow the forecasters to possess mutually correlated irrelevant signals. These choices are motivated by the one-off context where each forecaster provides only a single prediction, and thus the BIN model cannot be estimated. For instance, bias is a systematic error that is measured relative to the base rate of the outcome. Given that the base rate cannot be estimated without outcomes data, it is not possible to estimate bias in the ex-ante one-off context. Even though it is not possible to perfectly separate disagreement due to noise and information asymmetry based on one prediction per forecaster, extreme levels of disagreement are only possible in the presence of noise. This allows the RBA to estimate a lower bound on the level of noise and reduce the risk of overfitting.

   Satopää (2022) concluded by comparing a model that used the GJP data in order to benchmark RBA against other unsupervised aggregators, including those mentioned above. This horse race comparison was designed to illustrate the potential of RBA and it did not involve any empirical attempts to estimate the BIN pathways that might allow different aggregators to improve the accuracy of predictions. In addition to the predictions, RBA uses inputs comprising the common prior of forecasters, which is a probability prediction based on some of the forecasters' shared information. This is crucial for harnessing the forecasters' information and, indeed, several authors have argued that dispersed information cannot be merged without the forecasters' prior beliefs (e.g., Dietrich 2010). Unfortunately, forecasters in the GJP did not routinely report their prior beliefs. To include RBA in our comparison, we applied it with a default non-informative common prior of 0.5.[3] In our view, RBA provides a valuable contrast to averaging techniques because in addition to reducing noise, it has the potential to merge information dispersed among the forecasters.
6. PMs (Wolfers and Zitzewitz 2004). Forecasters act as traders who place bets on future events. A contract can pay $1 if the event happens, and $0 otherwise. If the current price is 60 cents, the supply–demand equilibrium in the market implies that the event has a probability of 0.6. If forecasters believe

[3] Satopää (2022) explained that at the beginning of a question when the uncertainty is highest and before forecasters have accumulated large amounts of question-specific evidence, they are likely to interpret a similar body of evidence, and hence disagree largely due to noise. In this study, the simple average of all probability predictions on the third day of a question was used as a "compromise" common prior. In section S2 of the Supplementary Material, we show that this simple average obtained very similar results compared to using the uniform prior of 0.5.

that the market has mispriced the chances, they can buy (sell) and hence increase (decrease) the price. The efficient market hypothesis posits that the current price should, in principle, reflect all publicly available information because as soon as forecasters become aware of noise, bias, or missing information, they have incentives to trade, thereby causing appropriate adjustments to the current price.
In classic micro-economic theory, PMs should be superior to averaging techniques (which only reduce noise) and RBA (which reduces noise and merges partial information). Indeed, in theory, markets should be the most efficient possible method for improving all BIN components. In addition this method does not rely on statistical aggregation but instead on crowds aggregating their own predictions, so it can provide insights into the advantages and disadvantages of statistical aggregation.

The six unsupervised aggregators described above are ordered roughly by the number of BIN components (bias, information, and noise) that they should theoretically improve, but this comes at a cost in terms of complexity. The average or median probability only requires a back-of-a-napkin type calculation. RBA depends on a sophisticated numerical Bayesian procedure that considers all potential noise/information asymmetry splits and finds an aggregate by weighting all possible scenarios by their respective likelihoods. Finally, PMs require a trading platform and liquidity management.

Supervised aggregators: Supervised aggregators depend on unknown variables that must be inferred from historical performance data. We determined which of the forecasters was closest to being correct about specific events by considering a range of supervised aggregators and evaluated them by performing leave-one-out cross validation. First, we formed a test set by separating one outcome and its predictions from the remaining data. Next, we estimated the unknown variables for the supervised aggregators using the remaining data (the training set). Finally, with our estimated model, we predicted the event in the test set by aggregating the predictions in the test set. Each aggregate prediction was then out-of-sample and it was trained based on equal but maximal amounts of training data, so our results were more likely to represent the inherent characteristics of the aggregation components.

We separated the supervised aggregators into "modules" that addressed different aspects of aggregation. Based on the study by Atanasov, Rescober, Stone, Swift, Servan-Schreiber, Tetlock, Ungar, and Mellers (2017), we focused on the effects of three modules controlled by five unknown variables denoted by ν_1, ν_2, ν_3, ν_4, and ν_5, as follows.

1. Temporal decay (TD): We calculated the probit average based only on a fraction $\nu_1 \in [\nu_{1,min}, 1]$ of the most recent predictions. To improve the stability of our estimation procedure, the lower bound $\nu_{1,min}$ was set such that each aggregate involved at least three predictions.
   We hypothesized that TD would increase the information levels of the aggregate because forecasters accumulate information over time. Thus, excluding older predictions will avoid diluting the aggregate with less informed predictions and instead give more weight to the most current and presumably informed predictions.
2. Differential weighting (DW): If we suppose that there are K forecasters, the equally weighted average would assign a weight $1/K$ to all predictions. We consider a variant of the probit average where the weights placed on forecasters' predictions depend on: (a) their historical accuracy and (b) how often they have updated their predictions regarding a question. A forecaster k receives a weight proportional to $w_{k,acc}^{\nu_2} \times w_{k,upd}^{\nu_3}$, where $w_{k,acc} \in [0, 1]$ is the kth forecaster's accuracy[4] at predicting the outcomes in our training set, $w_{k,upd} \in [0, 1]$ is proportional[5] to the number of times forecaster k updated the predictions of the test event, and $\nu_2, \nu_3 \in [0, 1]$ are parameters that determine how sensitive the aggregator is to heterogeneity in terms of accuracy and frequency during forecast updating, respectively.
   This approach could potentially obtain improvements through all three pathways. First, it assigns high weight to forecasters who performed well in previous prediction tasks. Clearly, good performance suggests that these forecasters had low levels of bias and noise, or larger amounts of information. Second, more updating requires effort and those who made greater effort to search for relevant signals in the news, as well as correcting biases and reducing noise were likely to be better forecasters (Atanasov et al., 2017).
3. Recalibration (RC): We transformed the forecasters' average probit prediction with a linear function. In particular, the recalibrated aggregate forecast is given by $\Phi(\nu_4 + \nu_5 \bar{P})$, where $\bar{P}$ denotes the crowd's (potentially weighted or decayed) probit average, $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution, ν_4 is the intercept of a linear transformation that makes the final aggregate systematically higher or lower, and ν_5 is the slope of the linear transformation that makes the final aggregate more or less extreme. On a calibration plot (recall Fig. 1), this mechanism can both shift and rotate the points and polarize the distribution of the predictions, thereby allowing RC to improve all three BIN dimensions.

[4] We measure accuracy as $1 - \text{BrS}_k$, where BrS_k is the forecaster's average Brier score in the training set. For instance, if the forecaster predicted 0.5 and 0.7 for two events that both occurred, then $1 - \text{BrS}_k = 1 - \frac{1}{2}[(1-0.5)^2 + (1-0.7)^2] = 1 - 0.17 = 0.83$. The scores range from 0 (perfect inaccuracy) to 1 (perfect accuracy). Each forecaster who had no past predictions in our training set received an accuracy score of 0.75, which represents the accuracy guaranteed by a constant (naïve) prediction of 0.5. In this manner, the forecasters' past predictions determined whether a forecaster was treated as net helpful (accuracy higher than 0.75) or net harmful (accuracy lower than 0.75).

[5] We counted how often forecasters updated their predictions for the test event and then normalized all counts by the maximum number of updates by each forecaster.

It should be noted that the parameters used in this approach could revert to the probit average if not supported by the training data. In addition to nesting the probit average, this approach treats the model as modular, where each of the three modules described above can be included or removed. For instance, one model could only recalibrate the probit average and not perform DW or TD, where the only module in this model was RC. Alternatively, we could recalibrate the probit average with DW in the model called RC + DW. Mixing and matching modules in this manner yielded seven different combinations of modules (TD, DW, RC, TD+DW, TD+RC, DW+RC, and TD+DW+RC), and thus seven aggregators. We studied the marginal effect of each module by training each of these aggregators.

6. Results

6.1. Unsupervised: Statistical aggregators

Our first analysis involved unsupervised aggregators, except for PMs, which we investigated later for the following two reasons. First, all unsupervised statistical aggregators, such as probability average and median, could be applied to forecasts across all four years of the tournament. Thus, we maximized the power with the statistical aggregators by using data from all four years. We then compared PMs using data from year 2, which had two markets with clear control groups of individual forecasters.

Second, given that the unsupervised aggregators can be applied to any single set of predictions, we could study their behavior in a diverse set of environments. Unfortunately, the same is not true of PMs. For instance, we cannot study PMs of different sizes based on past data. The market is a fixed entity that cannot be separated into pieces in this manner. However, we could apply the statistical aggregators to randomly sampled crowds with different sizes. In addition to crowd size, we evaluated the unsupervised statistical aggregators under different forecast time horizons, levels of information sharing among the crowd members, and levels of trimming and Winsorizing.

6.1.1. Level of trimming and Winsorizing

Jose et al. (2014) showed that trimming or Winsorizing forecasters' probabilities before averaging can boost their accuracy. To understand this effect, we applied different levels of trimming and Winsorizing to probability and probit predictions made 30 days before event resolution and averaged the remaining predictions.

Fig. 3 shows the improvements in the individual-level expected Brier scores (recall Fig. 2) due to aggregation in a 2×4 panel of plots. The columns correspond to all four combinations of trimming/Winsorizing and probability/probit average, as labeled at the top. Rows represent different experimental conditions for individual forecasters, as indicated by the labels on the left. The y-axis in each subplot shows the change in the expected Brier score. The vertical colored bars show how much the aggregators delivered improvements via bias, information, and noise. The x-axis in each subplot represents the level of trimming or Winsorizing, which increases from left to right.

The column on the far left in Fig. 3 represents the trimmed probability average. The top figure in this column shows that averaging probabilities with no trimming improved the expected Brier scores for untrained individuals by 0.07 through noise reduction. However, the probability average also increased the bias and reduced the level of information. The combined effect of the increased bias and decreased level of information led to a 0.02 increase in the expected Brier score. Therefore, the net improvement in the expected Brier score was about $0.07 - 0.02 = 0.05$. However, the negative effect of bias was almost eliminated by adequate trimming, as shown by moving right along the x-axis.

The effect of Winsorizing probabilities before averaging was similar to the effect of trimming, as shown in the second column from the left. The final two columns show the effects of trimming and Winsorizing probits before averaging them. As mentioned above, these mechanisms had very similar effects. However, in contrast to the probability average, excessive trimming or Winsorizing before averaging the probits for untrained individuals slightly increased the level of bias. The bottom row shows that trimming or Winsorizing had a negligible effect on the probability average and probit average of the predictions for superforecasters, thereby suggesting that trimming and Winsorizing were less effective for highly skilled forecasters.

However, in practice, there is no standard method for choosing the level of trimming or Winsorizing ex-ante based on a single set of predictions. Our results suggest that instead of trying to justify a particular level, the decision maker can rely on the unmodified probit average. Given that trimming and Winsorizing had very similar effects on the probability average, we only considered the trimmed probability average with 10% trimming, which we selected as a likely default choice.

6.1.2. Number of forecasters

Fig. 4 is similar to Fig. 3 except that instead of analyzing the effects of trimming or Winsorizing, we focused on unsupervised statistical aggregators and how they improved the accuracy of individuals (recall Fig. 2) based on predictions made by crowds of various sizes. The x-axis in each subplot represents the crowd size, which increased from five on the left to 75 on the right. Again, we fixed the forecast time horizon to 30 days before event resolution. To capture the trends for each size and aggregator, we formed 250 different crowds by randomly sampling the forecasters, calculated the accuracy improvement due to aggregation under each crowd, and plotted the average BIN decomposition.

These results demonstrate that averaging techniques (first four columns from the left) benefited from additional predictions when the crowd comprised untrained individuals. However, the improvement was rather small and mainly due to better noise reduction. Unfortunately, the larger crowd size did not help averaging to overcome the bias toward 0.5.

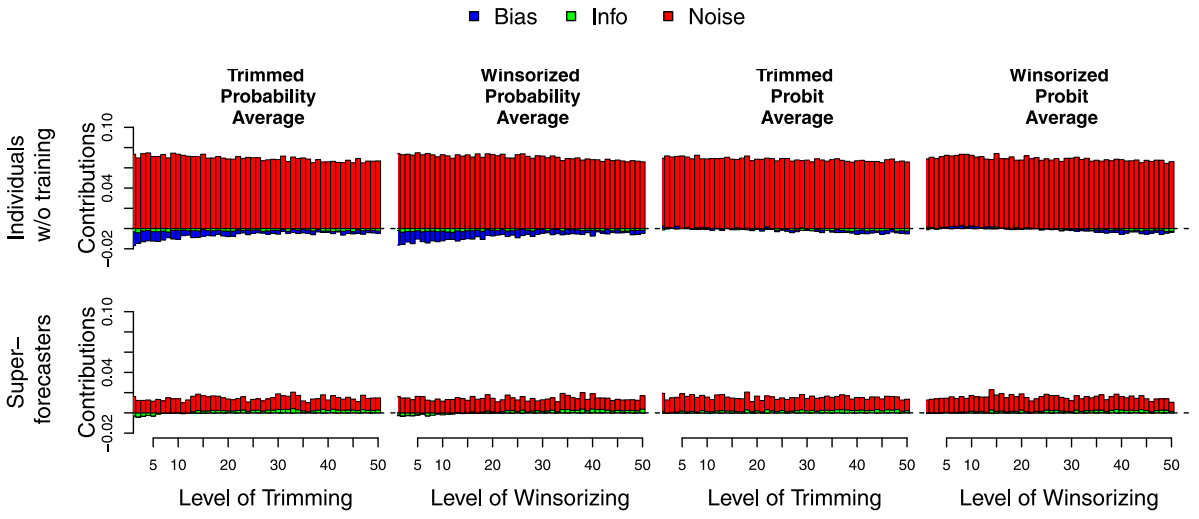

Fig. 3. Brier score contributions at different levels of (symmetric) trimming and Winsorizing of probability and probit predictions.

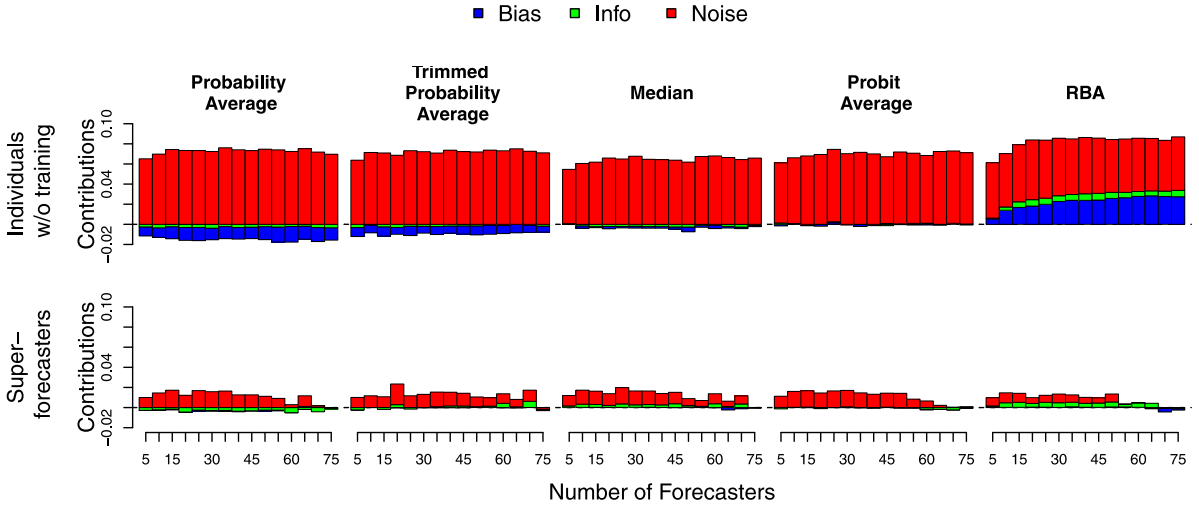

Fig. 4. Average Brier score contributions from unsupervised statistical aggregators under various sizes of crowds of forecasters.

By contrast, RBA obtained greater benefits with larger crowds of non-trained individuals, as shown by the column on the far right. Satopää (2022) stated that RBA assumes that larger crowds will possess more total information. Therefore, as the crowd size increases, RBA seeks to incorporate increasing amounts of information and obtains an extreme consensus prediction given by the probit average, which is very distant from the forecasters' common prior (0.5 in our case). Our results showed that the resulting aggregate was more informed and less biased than the typical individual forecaster. It was surprising that RBA could reduce the bias even though its model does not acknowledge bias in individual predictions. However, an aggregator does not need to model bias in order to reduce it. If individual forecasters are biased, an increasingly accurate aggregator must eventually reduce this bias. We consider this topic further in Section 6.1.4. However, making a large number of predictions is not necessary for RBA to boost the accuracy of individual forecasters. Indeed, even with only five predictions, RBA improved the accuracy as much as the best averaging technique, which was the probit average, and it achieved this by exploiting all three BIN dimensions.

6.1.3. Level of information sharing

Next, we considered crowds who differed in terms of the level of information shared among individual members. We sampled individuals who worked alone (with or without training) and placed them in the same crowd

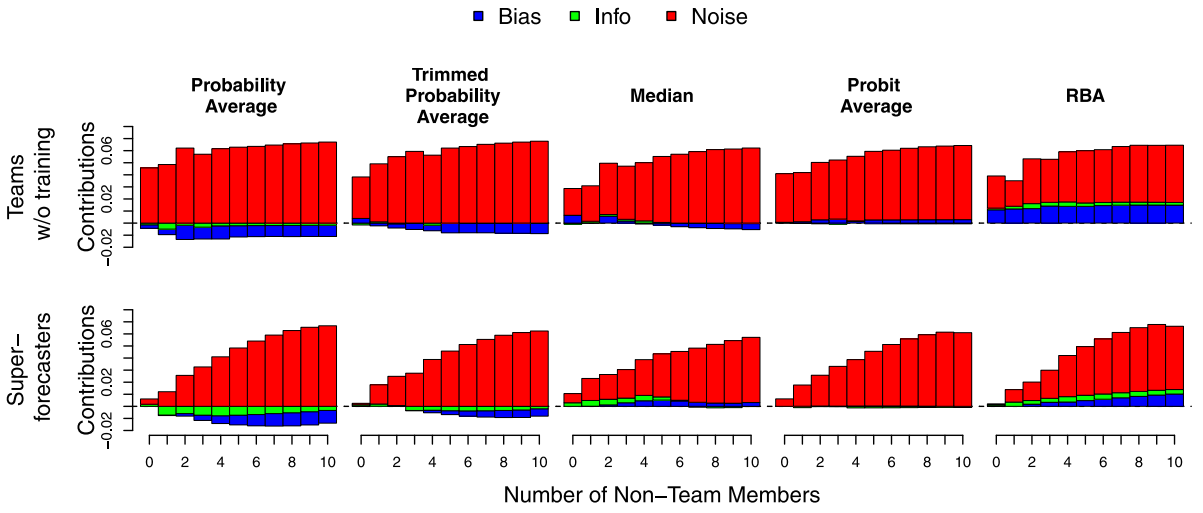

Fig. 5. Average Brier score contributions from unsupervised statistical aggregators under different levels of information sharing among crowd members.

with individuals who were known to work on the same team (in the teaming or superforecaster condition, respectively). For example, a 10-person crowd comprising three untrained forecasters who worked alone and seven untrained forecasters on the same team represented more information sharing than a crowd of seven untrained forecasters who worked alone and three untrained forecasters on the same team.

In our analysis, we fixed the forecast time horizon at 30 days, tested a crowd size of 10, and varied the number of non-team members from 0 to 10. To capture general trends, for each number of non-team members and aggregators, we formed 250 different crowds by augmenting randomly chosen teams with individual forecasters who worked alone, before calculating the accuracy improvement due to aggregation under each crowd and plotting the average BIN decomposition, as shown in Fig. 5. In Fig. 5, the *x*-axis in each subplot represents the number of non-team members, which increases from left to right. Therefore, crowds with more information sharing among members are shown on the left-hand side of each plot. To maintain a similar data set for each number of non-team members, we only considered questions that involved both non-trained individuals and teams. For the numbers of teams and questions available to each number of non-team members, please refer to section S1 in the Supplementary Material.

The results showed that when crowd members interacted with each other less, the RBA and probit average performed better compared with the other aggregators that lost accuracy due to increasing bias. Forecasters that worked alone were likely to make more different predictions, and thus the average became very close to 0.5.

Noise reduction became less important as the crowd members shared more information because the forecasters in teams had less noise (Satopää et al., 2021), and thus there was a lower capacity for noise reduction. This effect was particularly strong for RBA, which performed worse than probit average when the crowd consisted almost entirely of team members. In general, RBA seeks to model disagreement due to noise and information asymmetry in the crowd. However, in a one-off setting with limited data, it is not possible to make such distinctions perfectly. Therefore, in practice, RBA always assumes some level of both noise and information asymmetry. However, when forecasters can interact with each other, disagreements are likely to stem almost entirely from noise. RBA continues to assume that some of this disagreement stems from information asymmetry, thereby leading to overfitting, which adds noise, but RBA also performs noise reduction. The net effect of noise reduction and overfitting is still positive, thereby leading to less noise overall, and thus better accuracy. However, due to overfitting, RBA is less effective at reducing noise than simpler averaging techniques that assume all disagreement stems from noise, which is an assumption that aligns better with a crowd of team members. We consider overfitting further in our discussion of supervised aggregators.

6.1.4. Forecast time horizon

Fig. 6 decomposes the wisdom-of-the-crowd effect into the three BIN dimensions and shows how it changes as a function of the forecast time horizon, which increases from left to right. The shortest horizons (usually the easiest forecasting tasks) are shown on the left-hand side in each plot.

The top row in Fig. 6 shows how aggregators improved the accuracy of individuals without probabilistic training. Averaging techniques achieve this improvement almost entirely by noise reduction. The simple probability average increases the bias of longer-range predictions, but mainly because it pulls judgments toward 0.5 on the bounded probability scale, as explained by Baron et al. (2014). As illustrated earlier, trimming extreme predictions or using the median as a form of extreme trimming can alleviate these disadvantages. Probit averaging appeared to correct this undesirable tendency almost entirely.

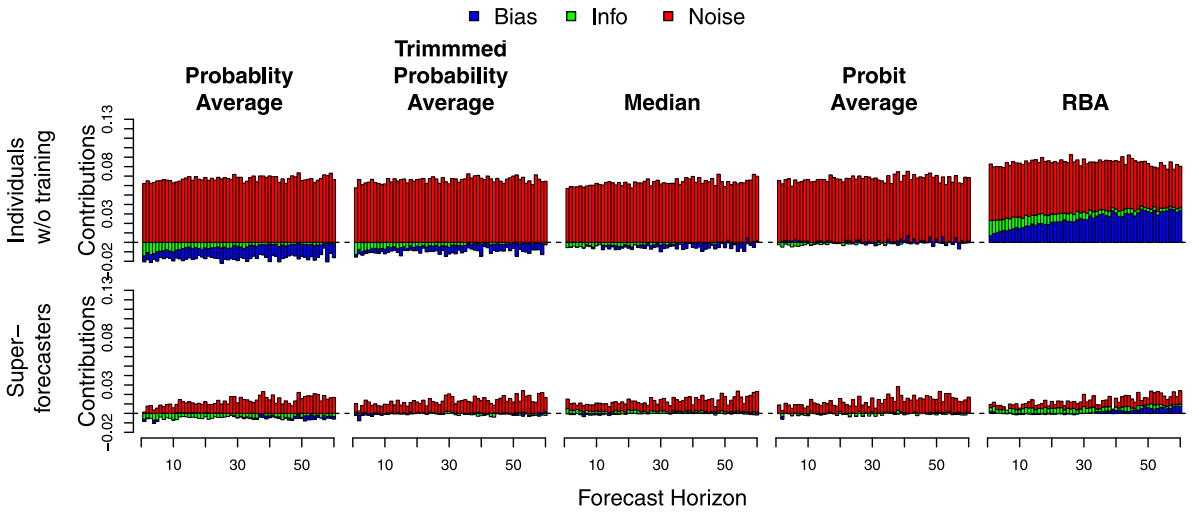

Fig. 6. Brier score contributions from unsupervised statistical aggregators under different forecast time horizons.

However, all of these aggregator are measures of central tendency, and thus they treat the disagreement of forecasters as noise. Thus, in principle, these aggregators may cause information loss whenever information asymmetry occurs. This effect was observed at short forecast time horizons (moving left in each subplot) and it decreased from left to right in the columns each subplot, and thus the probit average reduced the information level very little or none. Therefore, if decision makers want to reduce noise in predictions, they should consider probit averaging as an alternative to more popular measures of central tendency, such as the probability average, median, or trimmed probability average.

Overall, RBA was most successful at improving the accuracy, as shown by the height of the colored bars. Furthermore, unlike measures of central tendency that act almost exclusively through noise reduction, RBA delivered improvements across all three BIN dimensions: bias, noise, and information. About half of the improvement was due to noise reduction. For long forecast time horizons (toward the right in each subplot), RBA improved the accuracy through bias reduction more than by merging the dispersed information of individual forecasters. However, this ordering reversed as the horizon shortened (toward the left-hand side of the plots).

To understand this trend, we recall that bias reduction is only possible if forecasters are biased and information fusion is possible only if forecasters use dispersed sets of information. At longer horizons, forecasters are unlikely to draw on dispersed information. When the event resolution time becomes close, more news surfaces and attentive forecasters can create information asymmetries in the crowd, which can then be leveraged by the aggregators. To examine the nature of the forecasters' bias in our study, we considered the probit average. The probit average entirely affects the accuracy through noise reduction and preserves the level of information and bias for individual forecasters. In contrast to the probability average, which is often excessively close to 0.5, RBA aims to integrate the dispersed information of forecasters and it typically[6] forces the probit average to an extreme away from the user-specified common prior, which we set to 0.5 for non-informative prediction. Our results showed that forcing to this extreme reduced the forecasters' bias, particularly for long forecast time horizons. Therefore, although the probability average was systematically excessively close to 0.5, forcing the individual forecasts and their predictions to extreme values away from 0.5 could reduce this bias.

Finally, the bottom row in Fig. 6 shows how aggregators improved the accuracy of superforecasters. Averaging performed better when the individuals were highly skilled. For instance, superforecasters may have been approaching the limits of epistemic uncertainty so little further public information could be incorporated into their forecasts and the remaining disagreements were probably due to noise, which could be reduced by averaging. Our results suggest that the best approach was to average the predictions of superforecasters. The differences among aggregators were small and not statistically significant, as shown in section S3.4.1 of the Supplementary Material.

6.2. Unsupervised: Prediction markets

6.2.1. Forecast time horizon

During the second year of the tournament, the GJP explored two continuous double auction PMs run by Lumenogic, where the participants worked independently. In one market, they received training in how to maximize profits and avoid market errors. By contrast, there was no

[6] If the crowd is small (e.g., fewer than 20 individuals) and the forecasters' level of noise is high, RBA can shrink the probit average toward the common prior. However, in our analysis, the crowd size was typically larger than 50 (see the Supplementary Material). For these large crowds, the level of noise must be extremely high before the RBA shrinks the probit average toward the common prior. For more information, see Figure 4 in the study by Satopää (2022).

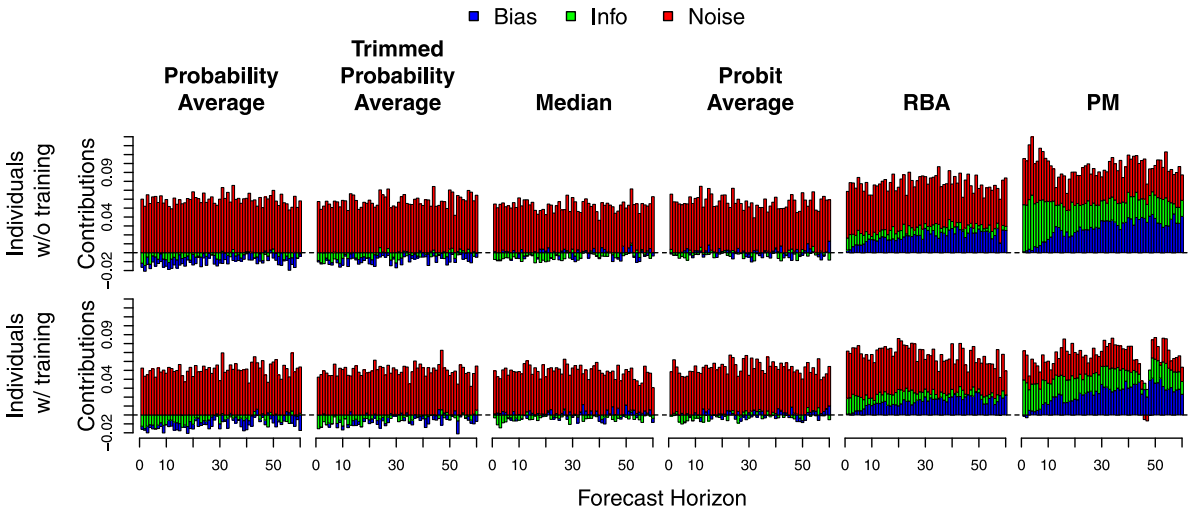

Fig. 7. Brier score contributions from unsupervised aggregators, including PMs, under different forecast time horizons.

training in the other market.[7] Therefore, we only considered individuals with and without training as the control experimental conditions.

Similar to the previous results, Fig. 7 shows the improvements in the individual level expected Brier scores due to aggregation at different forecast time horizons, and it illustrates the improvements due to bias reduction, noise reduction, and information acquisition. Given that the analysis used much less data than those shown in Figs. 3–6, the variability is greater between consecutive colored bars and the plots are more fluctuating.

However, for the unsupervised statistical aggregators, the results agreed qualitatively with those presented in the previous section, where averaging improved the accuracy almost entirely through noise reduction whereas RBA obtained improvement through all three BIN dimensions. The novel results are shown in the PMs column on the right, which indicates that PMs improved the accuracy through all BIN dimensions. Furthermore, PMs clearly improved the bias and information levels more than RBA, where PMs led to the highest reduction in the individual level expected Brier score among all of the unsupervised aggregators considered in this study.

6.3. Supervised: Statistical aggregators

6.3.1. Forecast time horizon

Next, we analyzed supervised aggregators that require past performance and outcome data. In practice, statistical models are rarely built once and the best models are tailored to specific problems. It is often sensible to start with a simple model, before indentifying problems and searching for incremental improvements (Draper, 1995; Schad, Betancourt, & Vasishth, 2020; Smith, 1984). Thus, we propose a modular aggregation construction method, which involves testing different modules and finding the best combination. This approach gives the researcher a more detailed view of new aggregators. By decomposing the aggregator, we can inspect each module, test ways to improve it, or spot redundancies.

As an analogy, we consider the construction of a racing car. The top speed of your model is similar to that of the models made by other teams, but poor acceleration is preventing you from reaching the podium. The braking system is unlikely to improve acceleration, so your team decides not to prioritize brakes, but instead it prioritizes the engine. Our modular construction of the aggregator allows a similar understanding of the modules. For instance, we may find that DW improves the accuracy through noise reduction. However, there are many ways to perform weighting. According to our analysis, we can then make targeted changes and search for the desired effect in a single module. We may find that the aggregator improves the overall accuracy of forecasters, but one module limits the potential and removing that module allows the aggregator to perform even better. We can also inspect interactions between modules. We may find that some modules work well or poorly together. The bottom line is that modular construction means that the development of aggregators is more systematic and robust.

To illustrate this idea, we considered all possible combinations of the three modules described in Section 5. In a similar manner to the results presented in Figs. 6–7 for unsupervised aggregators, Fig. 8 shows the results for supervised aggregators, where the accuracy improvements are decomposed into bias reduction, noise reduction, and enhanced extraction of dispersed information for forecasters over different forecast time horizons. The seven columns represent all possible combinations of the three modules.

The three columns on the far left in Fig. 8 correspond to models with individual modules (TD, DW, and RC).

[7] The third year included several predictions markets also run by Lumenogic. Instead of separating trained and untrained individuals into separate markets, all individuals participated in the same market, which made it impossible to separate the effects of training from other factors. Therefore, for simplicity, we focused only on data from the second year of the tournament when we had a clear control group for each PM.

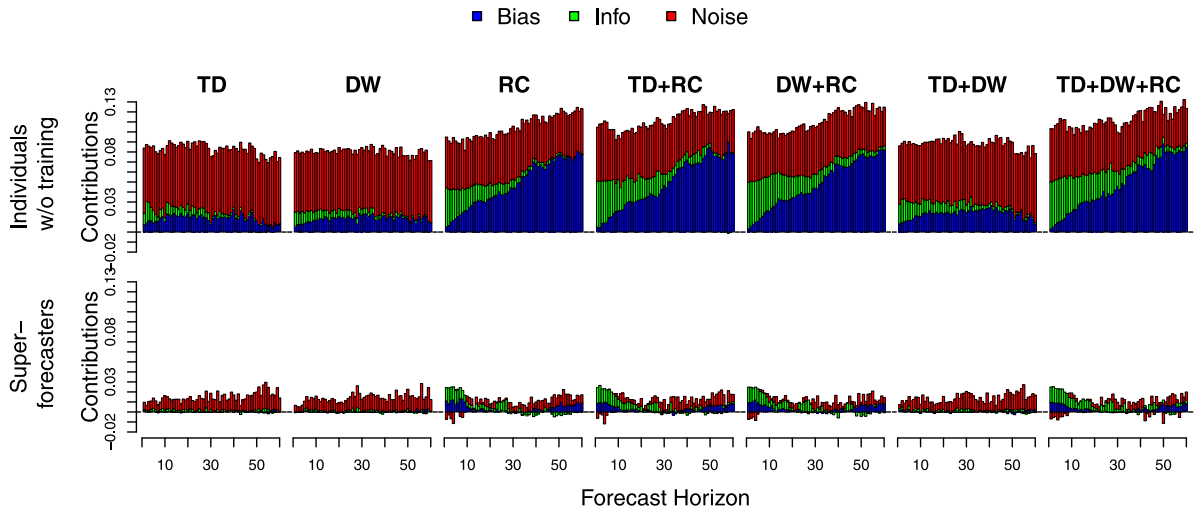

Fig. 8. Brier score contributions from supervised aggregators with different combinations of the modules over different forecast time horizons.

All three modules improved the accuracy through noise reduction. Each module exploited the probit average as an effective method for reducing noise in individual predictions. Comparisons of the probit average results in Fig. 6 with the TD and DW results in Fig. 8 show that these modifications to the probit average could reduce the bias and combine the dispersed information of forecasters. Given that both TD and DW are weighted averages, they could have achieved these improvements by assigning greater weights to forecasters who were expected to have lower bias and more information. The only clear change from TD and DW compared with TD+DW was the slightly increased level of information in TD+DW, which suggests that TD and DW captured slightly different parts of the dispersed information of forecasters.

However, measures of central tendency such as the weighted averages that drive TD and DW are limited by the extent to which they can shift predictions systematically up or down, which could correct a bias or make predictions more extreme and increase the level of information (recall Fig. 1). By contrast, RC is less limited, and thus it was more effective at bias reduction and information fusion than TD and DW. In addition to the slightly increased level of information in the more complex models, the single module RC model performed as well as or better than the more complex models with two or three modules (results shown in the four columns on the right-hand side in Fig. 8).

As once stated by Albert Einstein: "Everything should be made as simple as possible, but not simpler". Based on this principle, which is essentially a refinement of Occam's razor, we should search for the simplest model but without compromising the accuracy excessively. TD and DW captured slightly different parts of the forecasters' information, but the marginal benefit in terms of accuracy when including both modules might not be worth the added computational burden incurred to estimate their parameters. Thus, including both TD and DW will lead to redundancy. A decision maker may prefer the simpler three-parameter TD+RC to the four-parameter DW+RC or five-parameter TD+DW+RC, which obtained similar levels of accuracy.

This analysis is illuminating but it does not provide insights into how the modules interact and contribute to accuracy improvements. To understand these interactions, we can generalize the BIN decomposition given by Satopää et al. (2021) by further splitting the three BIN modules into the effects due to RC, TD, and DW. Fig. 9 focuses on the column on the far right in Fig. 8 (TD+RC+DW) and shows how each module contributes to the accuracy of this three-module model, where from left to right, the three columns correspond to the contributions from bias, information, and noise. The rows represent the different experimental conditions for individual forecasters and they are labeled on the left. The colored bars show how each module changes the individual level expected Brier score through each BIN dimension. For instance, the decomposition of bias in the column on the far left essentially shows that all bias reduction is due to RC and that this contribution decreases along the forecast time horizon. Similarly, the middle column shows that in the three-module model, most of the information acquisition occurs via RC. The column on the far right shows that the task of noise reduction in the aggregator is split almost equally between TD and DW.

Interestingly, RC could introduce additional noise over long forecast time horizons (toward the right-hand side in each plot in the column on the far right). In section S4.1.2 of the Supplementary Material, we show that this increase in noise was particularly pronounced when we aggregated predictions from untrained teams, possibly because we had the least data under this condition. One explanation for this pattern is overfitting, where the RC mechanism might have falsely interpreted irrelevant signals as information, which demonstrates how merging the information of forecasters can be a double-edged sword. Perfect fusion requires that the aggregator can distinguish two types of variability comprising variability that covaries with outcomes (information) and variability that

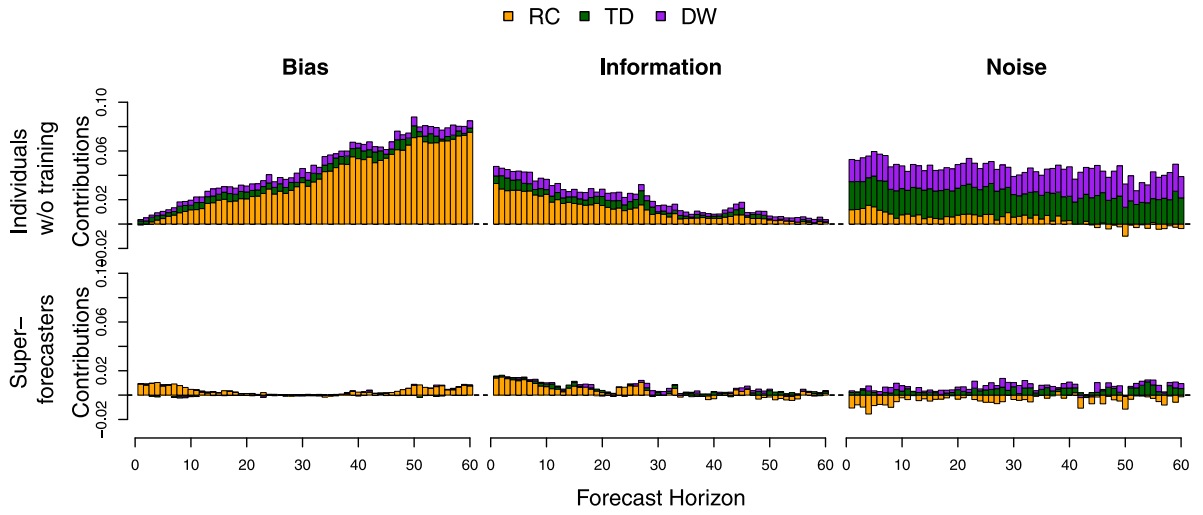

Fig. 9. Detailed decomposition of the improvements attributable to the different modules in the full TD+DW+RC model due to bias reduction, information fusion, and noise reduction.

does not (noise), and any errors will lead to overfitting. By contrast, averaging is less ambitious because it does not aim to merge dispersed information, but instead it applies a more conservative approach of treating all variability as noise, which avoids overfitting but at the cost of excessive conservativeness. Winkler et al. (2019) made a similar distinction and they explained that sophisticated techniques can produce more accurate predictions but they are prone to overfitting, and thus they are riskier than simple rules such as averaging.

Returning to our racing car analogy, Fig. 9 suggests a method for improving the three-module model, where we could make the RC module robust against overfitting by regularizing the estimation of the RC parameters ν_4 and ν_5 with elastic net or other related techniques (Zou & Hastie, 2005). However, implementing these approaches is more complicated and outside the scope of the current study. It should be noted that this direction for improvement is only visible in Fig. 9 and could not have been deduced from the less granular analysis presented in Fig. 8.

7. General discussion

The results obtained in the present study demonstrate the value of moving beyond horse racing comparisons focused on the aggregators that best reduce the error scores, but instead identifying the pathways that allow promising methods to improve the accuracy of predictions (Davis-Stober et al., 2014; Soll & Larrick, 2009). Our empirical findings generally agreed with our theoretical expectations, except for those mentioned in the following.

As expected, averaging techniques (unweighted or weighted) boosted the accuracy almost entirely by noise reduction. Simple averaging also increased the bias mainly because it can pull the forecasts toward 0.5 on the bounded probability scale (Baron et al., 2014). Trimming extreme probabilities before averaging can alleviate this bias, but in practice, there is no standard method for selecting the level of trimming in an ex-ante one-off forecasting context. However, our results suggest that the probit average provides a simple alternative that can correct this tendency without needing to select the values of tuning parameters. All of the aggregators produce measures of central tendency and they may theoretically reduce the level of information because they treat all disagreement as noise, thereby eliminating disagreements that arise from different sets of information. Our empirical analysis confirmed that this effect applied to all of the averaging techniques except for the probit average, which generally preserved the forecasters' level of bias and information, and improved the accuracy of the predictions only through noise reduction.

Thus, greater bias reduction and fusion of information offer opportunities for further improvement. However, exploiting these opportunities will incur costs in terms of increased complexity and data requirements. In particular, Dietrich (2010) stated that combining the dispersed information of forecasters with a statistical aggregator is not possible without their prior beliefs. As a novel methodology, Satopää (2022) devised an RBA that is estimated with numerical Markov chain Monte Carlo techniques and the common prior and predictions of forecasters as inputs. The GJP did not collect the prior beliefs of forecasters but we demonstrated how using a non-informative uniform prior of 0.5 still allowed RBA to improve the accuracy by exploiting all three BIN dimensions. We consider that further improvements could be achieved by using the actual prior beliefs of forecasters.

Overall, our findings suggest that by introducing more rigor and discipline, the state-of-the-art practice of crowd wisdom utilization can be extended well beyond mere averaging and noise reduction. If we only continue eliciting the predictions of forecasters, statistical aggregation will remain focused on noise reduction. Thus, it is time to consider a more principled process where forecasters first search for an appropriate base rate and then update it to consider their context-specific information. This approach demands more effort but it can reduce bias by centering

the individual predictions more appropriately, before then allowing aggregators such as RBA to reduce the noise and merge the dispersed information of forecasters.

However, RBA did not enhance the information extraction process as well as the PMs. PMs are much more complicated and often expensive to implement than averaging or RBA but they can also provide a broader range of benefits. As expected, they boosted the accuracy via all three components of the BIN model. PMs are often well calibrated (Atanasov et al., 2017), which suggests that they reduce bias and noise. Among the aggregators that we considered, PMs performed best at improving the extraction of (partial) information. This result is consistent with micro-economic arguments about the power of markets for facilitating effective information sharing among individuals. It is possible that information aggregation is best handled by market participants rather than mechanical aggregators, at least in the context considered in the present study. However, statistical aggregation is slightly more efficient at reducing bias and noise. Humans who work in markets may be better at extracting information in geopolitical tournaments and machines might be better at reducing noise and bias (Atanasov et al., 2017).

These findings suggest a strategy that allows decision makers to combine the dispersed information of forecasters without using complex aggregators such as RBA or PMs, where the idea involves altering the forecasters' level of bias, information asymmetry, and noise to better match averaging. For instance, allowing forecasters to exchange information in teams will make their predictions more amenable to averaging. In a case study of an anonymized electronics company, Oliva and Watson (2009) showed how demand forecasting can be improved by forming a team dedicated to collecting and organizing relevant information from different functional units of the company (capturing the data nominations from finance, operations, marketing, and so on). The team then asks all units to make their predictions based on this common information. This information management strategy reduces information asymmetry and should boost the performance of averaging. Alternatively, the decision maker can conduct a noise audit (Kahneman, Sibony, & Sunstein, 2021) that directs forecasters to decompose the signals they are using for making probability judgments, before independently assessing their values and then weighting cues accordingly. For instance, forecasters might make different probability judgments because they perceive different base rates for an event or because they see the event as more similar to one historical analogy than another. Exposing these disagreements can identify errors and after adjusting their predictions, forecasters might be able to reconcile their disagreements, thereby reducing both noise and information asymmetry. In brief, if decision makers choose to average predictions, they should match this decision with a management process that gives forecasters access to a common set of information.

In general, when there is less epistemic uncertainty, information fusion is less important and averaging is more appropriate. Superforecasters who operate in the rather noisy domain of geopolitics might even be approaching the limits of epistemic uncertainty. Initially, it may appear plausible that a superforecaster has acted strategically and chosen to not share some private information with the others. However, according to our results, the best approach involved averaging the predictions of superforecasters. Except over short forecast time horizons, averaging the predictions of superforecasters performed as well as even the supervised techniques with access to past performance and outcomes data. Therefore, most if not all of the superforecasters' disagreement was likely to stem from noise, thereby suggesting that very little information was actually held back and that little more useful public information could be incorporated into their forecasts.

Over different forecast time horizons, noise reduction remained the most important contributor to the wisdom of the crowd effect. We suspect that long-horizon forecasts benefited more from bias reduction than information fusion because there was less information asymmetry among forecasters. For example, few forecasters possessed insider information about how the Syrian civil war would end when it started in 2011 or how a US Presidential election would end before the primaries had begun. Forecasters are often roughly equally ignorant about far-off futures. Thus, most of the gains come from bias or noise reduction, where statistical aggregators are most helpful. These findings suggest a strategy for making long-range predictions, where bias can be reduced by providing the forecasters with relevant base rates, before then reducing noise by averaging their (probit) predictions. If the use of base rates can successfully remove bias and yield longer-range predictions that disagree mainly due to noise, then averaging can outperform RBA, which always assumes some degree of information diversity and it can result in overfitting. For shorter-range predictions, information asymmetry emerges when relevant information is more available. According to our discussion above, the decision maker may search for ways to give forecasters access to a common set of information before requesting and averaging their predictions. However, if this is not possible and the decision maker is left with predictions that disagree due to information asymmetry, then information fusion techniques such as RBA are likely to yield better results than averaging.

This raises the following question: over what forecast time horizon should one switch from averaging to an approach such as RBA? The correct timing depends on when relevant news begin to emerge, and thus it is likely to depend on the context. Machine learning techniques can help to track the evolution of news (e.g., Allan, Carbonell, Doddington, Yamron, and Yang 1998). Alternatively, in an ex-ante context without outcomes data, it may be possible to use natural language processing techniques developed by computational linguists and analyze the written rationales behind the predictions of forecasters (Karvetski, Meinel, Maxwell, Lu, Mellers, & Tetlock, 2021). Implementing a method for inferring the level of information asymmetry and noise from the written rationales of forecasters is a challenge that we reserve for future work.

These examples illustrate the complementary benefits of aggregation and elicitation. However, these benefits

extend beyond matching aggregators to environments. Knowing how an aggregator boosts the accuracy of predictions enhances our understanding of the mechanism involved, thereby providing guidance to facilitate further methodological improvements. Based on BIN decomposition, we propose the modular construction of supervised aggregators, which provides a granular view of each module in the aggregator and improvements can be identified by highlighting potential redundancies or shortcomings. A similar approach to an experimental intervention could be applied to improve individual forecasters. For instance, we might consider designing a training program or elicitation scheme that involves calibration training, outcome feedback, and team discussion. However, it may be impractical to incorporate all modules due to participant fatigue, and thus some choices must be made. By eliciting predictions under each combination of modules, our method suggests approaches for mixing and matching modules to maximize improvements across three dimensions within the constraints of the participants. Clearly, the total number of combinations becomes huge with many modules, thereby making full factorial designs impractical. However, a partial solution is a fractional factorial design that avoids redundancies by carefully selecting subsets of combinations from the full factorial design (e.g., see Montgomery 2017).

The modular construction of aggregators is not affected by the same data collection challenges. We can simply collect predictions and apply all possible combinations of aggregation modules to the data, as described in Section 6.3. The analysis can reveal redundancies, prune unnecessary parameters, and allow the development of unsupervised aggregators in one-off contexts where data are often limited. In these contexts, the model must balance the important aspects of the problem with what can be estimated from the data (Palley & Satopää, 2022). At present, the actual modular construction of a one-off aggregator is unclear but it is an intriguing challenge for future research.

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.

Acknowledgments

This work was supported by the INSEAD-Wharton Alliance and the Intelligence Advanced Research Projects Activity (IARPA), contract number 140D0419C0049. The U.S. Government is authorized to distribute reprints notwithstanding any copyright annotation. The views expressed here are those of the authors, not of IARPA or the U.S. Government.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2021.12.010.

References

Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (1998). *Topic detection and tracking pilot study final report*. Carnegie Mellon University.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners, Vol. 30*. Springer Science & Business Media.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., et al. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

Chen, Y., & Pennock, D. M. (2010). Designing markets for prediction. *AI Magazine*, *31*(4), 42–52.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*(2), 187–203.

Cross, D., Ramos, J., Mellers, B., Tetlock, P. E., & Scott, D. W. (2018). Robust forecast aggregation: Fourier L2E regression. *Journal of Forecasting*, *37*(3), 259–268.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79.

Dietrich, F. (2010). Bayeslan group belief. *Social Choice and Welfare*, *35*(4), 595–626.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *57*(1), 45–70.

Elliott, G., & Timmermann, A. (2013). *Handbook of economic forecasting*. Elsevier.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519.

Hora, S. C., Fransen, B. R., Hawkins, N., & Susel, I. (2013). Median aggregation of distribution functions. *Decision Analysis*, *10*(4), 279–291.

Jose, V. R. R., Grushka-Cockayne, Y., & Lichtendahl Jr, K. C. (2014). Trimmed opinion pools and the crowd's calibration problem. *Management Science*, *60*(2), 463–475.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: a flaw in human judgment*. New York: Farrar, Straus & Giroux.

Karvetski, C., Meinel, C., Maxwell, D., Lu, Y., Mellers, B., & Tetlock, P. E. (2021). Forecasting the accuracy of forecasters from properties of forecasting rationales. Working Paper of Good Judgment 2.0 in IARPA FOCUS program.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS One*, *13*(3), Article e0194889.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276.

McAndrew, T., Wattanachit, N., Gibson, G. C., & Reich, N. G. (2021). Aggregating predictions from experts: A review of statistical methods, experiments, and applications. *WIREs Computational Statistics*, *13*(2), Article e1514.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., et al. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267–281.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: eliciting experts' probabilities*. Chichester: John Wiley & Sons.

Oliva, R., & Watson, N. (2009). Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. *Production and Operations Management*, *18*(2), 138–151.

Palley, A. B., & Satopää, V. A. (2022). Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. Available at SSRN: https://ssrn.com/abstract=3504286.

Palley, A. B., & Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, *65*(5), 2291–2309.

Satopää, V. A. (2021a). Improving the wisdom of crowds with analysis of variance of predictions of related outcomes. *International Journal of Forecasting*, *37*, 1728–1747.

Satopää, V. A. (2022). Regularized aggregation of one-off probability predictions. *Operations Research*, http://dx.doi.org/10.1287/opre.2021.2224.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356.

Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, *111*(516), 1623–1633.

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. *Management Science*, *67*(12), 7599–7618.

Schad, D. J., Betancourt, M., & Vasishth, S. (2020). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, Available at arXiv: https://arxiv.org/abs/1904.12765.

Smith, A. F. (1984). Present position and potential developments: Some personal views bayesian statistics. *Journal of the Royal Statistical Society: Series A (General)*, *147*(2), 245–257.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions.. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780.

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

Sunstein, C. R., & Hastie, R. (2015). *Wiser: getting beyond groupthink to make groups smarter*. Harvard Business Press.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: the art and science of prediction*. Random House.

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl Jr, K. C., & Jose, V. R. R. (2019). Probability forecasts and their combination: A research perspective. *Decision Analysis*, *16*(4), 239–260.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*(2), 107–126.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320.