Decomposing the Effects of Crowd-Wisdom Aggregators: The Bias-Information-Noise (BIN) Model

Ville A. Satopää

INSEAD, Fontainebleau, France

Marat Salikhov

Yale School of Management, USA

Philip E. Tetlock, Barbara Mellers

The Wharton School of the University of Pennsylvania, USA

Abstract

Aggregating predictions from multiple judges often yields more accurate predictions than relying on a single judge: the wisdom-of-the-crowd effect. But there is a wide range of aggregation methods, from one-size-fits-all techniques, such as simple averaging, prediction markets, and Bayesian aggregators to customized (supervised) techniques, such as weighted averaging, that require past performance data. This article applies a wide range of aggregation methods to subjective probability estimates from geopolitical forecasting tournaments. It uses the Bias-Information-Noise (BIN) model to disentangle three mechanisms by which aggregators improve accuracy: the tamping down of bias and noise and the extraction of valid information across forecasters. Simple averaging works almost entirely by reducing noise, whereas more complex techniques, like prediction markets and Bayesian aggregators, work via all three pathways: better signal extraction as well as noise and bias reduction. We close by exploring the utility of a BIN approach to the modular construction of aggregators.

Keywords: Judgmental Forecasting, Partial Information, Prediction Markets, Wisdom of Crowds

1. Introduction

A vast research literature attests to the power of aggregating predictions to improve accuracy (for excellent reviews, see Clemen 1989; Clemen and Winkler 1999; Armstrong 2001; Winkler et al. 2019). This literature explores a wide range of methods of aggregating predictions, from machine learning and statistical techniques (Makridakis et al., 2018; McAndrew et al., 2021) to prediction markets (Chen and Pennock, 2010; Elliott and Timmermann, 2013). This literature also examines ways of selecting and assessing forecasters (Mannes

[1]Corresponding author. E-mail: ville.satopaa@insead.edu. Supported by the INSEAD-Wharton Alliance and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0419C0049. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. Disclaimer: The conclusions contained herein are those of the authors and should not be interpreted as those of IARPA, DOI/IBC, or the U.S. Government.

et al., 2014; Budescu and Chen, 2015), of organizing interactions (Tetlock and Gardner, 2016), and of sharing information (Sunstein and Hastie, 2015; Palley and Soll, 2019; Satopää et al., 2016). In numerous fields, there is growing interest in moving beyond horse-race comparisons of which aggregator shrinks error scores the most to identifying the mechanisms by which promising methods improve accuracy (Erev et al., 1994; Davis-Stober et al., 2014; Soll and Larrick, 2009).

Recently, Satopää et al. (2021) proposed a new statistical framework, the Bias-Information-Noise (BIN) model, for explaining how experimental treatments in forecasting tournaments can improve individual forecasters' accuracy 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 effects of tournament interventions such as training in probabilistic reasoning, working in teams, and tracking talented forecasters then putting them in elite teams.

Satopää et al. (2021) focused, however, solely on modeling the treatment effects on individual forecasters, not on strengths and weaknesses of aggregators. By contrast, our focus is exclusively on aggregators. By analyzing aggregation techniques in the same way that Satopää et al. (2021) analyzed experimental treatments, we can move beyond tabulating outcomes of horse-races to a deeper understanding of aggregators. A more apt metaphor becomes professional car races, where mechanics can peak under the hood and inspect the performance engines underlying aggregators. We apply a range of aggregators to a series of probability predictions in forecasting environments that vary in terms of the size of the crowd, the level of skill and information sharing within the crowd, and the forecast time horizon. Our goal is to understand the pathways by which different types of aggregators synthesize wisdom from crowds.

Such granularity is important for two reasons. First, predictions elicited in different environments are likely to have different levels of bias, noise, and information asymmetry. Our choices of aggregation tools should match the environment. For instance, we pay a potentially steep accuracy price if we apply a pure noise-reduction aggregator to predictions that vary due to information asymmetry (different forecasters possessing distinct pockets of valid knowledge). Second, we can use the BIN decomposition to advance the modular construction of aggregators, which involves partitioning an aggregator into modules, inspecting how each module contributes to overall accuracy, and detecting redundancies and potential for improvements.

2. Bias, Information, and Noise

The BIN model, advanced by Satopää et al. (2021), is the analytic engine that traces variation 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. It treats the signals as causal: the target event occurs if and only if the cumulative signal contribution is positive. In addition to relevant signals, the universe contains signals of no relevance to the event. Forecasters sample and interpret all signals with varying skill and thoroughness.

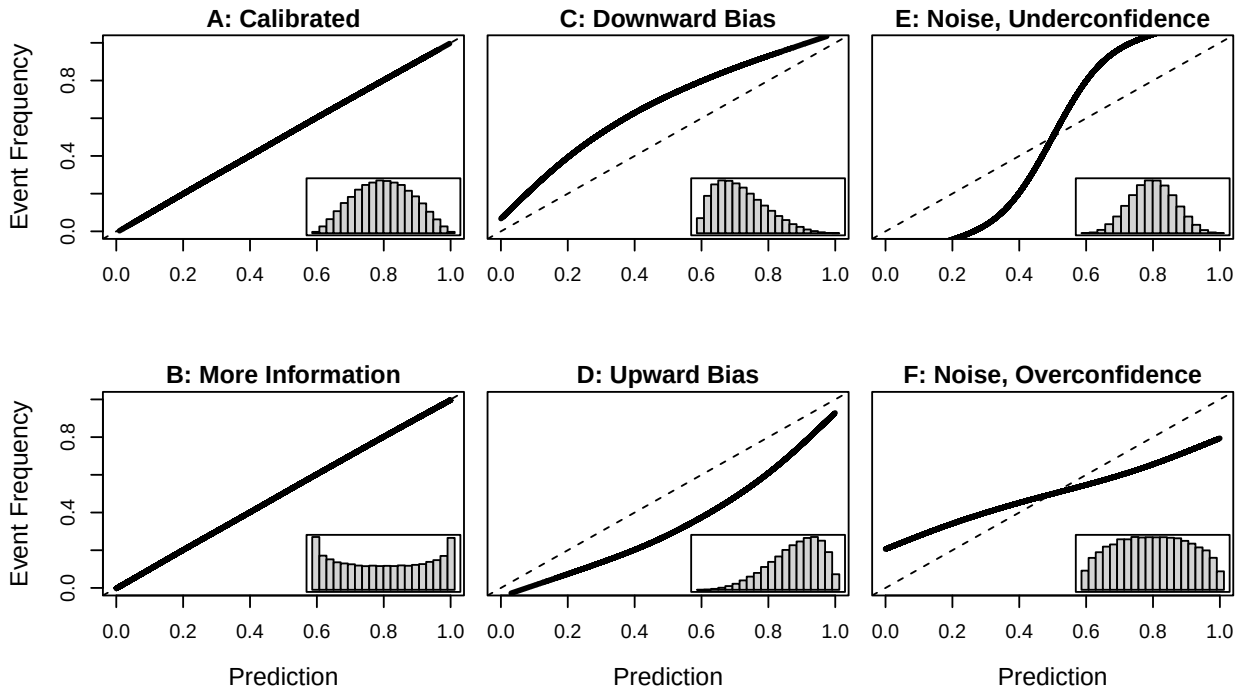

Figure 1: From left to right, columns show the effects of information, bias, and noise, respectively. The inlaid figures in the bottom-right corners describe the distributions of predictions. Each distribution has its own x-axis defined by its bottom border and ranges from 0 (left) to 1 (right).

Their samples consist of relevant signals (partial information) and irrelevant signals (noise). The BIN model makes the bounded rationality assumption that forecasters have 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, producing systematic biases. For a stylized example, see Example 1 in Satopää et al. (2021).

Each pathway has a distinctive observable effect on forecasters' predictions. To illustrate, we use the BIN model to simulate outcomes with a base rate 0.5 and forecasters' predictions with different levels of bias, noise, and information. Figure 1 uses calibration plots to graph these predictions against the objective frequency of event occurrence. To track how extreme or confident the predictions are, the inlaid boxes in the bottom right corners contain histograms of the predictions that the forecaster would make for a large number of events.

Panel A depicts a perfectly calibrated forecaster, one without bias or noise. The calibrated forecaster's probability predictions align perfectly with the objective frequencies of events. Whenever the forecaster reported a probability of, say, 0.2, exactly 20% of the events happened; and the equivalent relation holds for all possible predictions between 0 and 1 – not just 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, predictions still align perfectly with the objective frequencies

of events. However, predictions in panel B are more extreme or decisive, showing greater resolution than those in panel A. The panel-B forecaster would be more valuable to decision makers because that forecaster has all three BIN advantages: 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 has a downward bias; panel D, 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, making the forecasters' predictions more extreme than can be justified. In panel E, on the other hand, irrelevant signals are negatively correlated with the forecaster's relevant signals, which leads to predictions that are too close to the non-informative base rate 0.5 and hence are under-confident. In panels C, D, E and F, predictions no longer align with objective frequencies of events, 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

This section briefly reviews 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^* that represents 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 the expected levels of bias, 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 that represents 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 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 forecasters' accumulations of signals to probability predictions, while ensuring 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 believe that $\delta = 0$ and $\mu = 0$. With the potential for misbeliefs, the forecaster predicts:

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

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

In this article, we model the aggregator as the individual predictions above, except that the aggregator is

treated as a group of its own.[2] As with individual predictions, the aggregate forecast is based on a normally distributed variable that represents the accumulation of all signals revealed to the aggregator by individual forecasters' predictions. This variable is then converted into a probability prediction under the same bounded rationality assumption as 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 individual forecasters' expected accuracy and perform a marginal analysis that explains performance differentials in terms of reduction of noise, tamping down of bias, or effective use of dispersed information in crowd predictions. In this way, we can explore the pathways by which aggregators derive wisdom from crowds. In the next sections, we conduct such an investigation drawing on series of subjective probability predictions made in different forecasting environments.

4. Individual Forecasters and Events

We use the BIN model to analyze aggregators in a 4-year series of probability estimates that forecasters made in geopolitical tournaments sponsored by the research branch of the U.S. intelligence community: IARPA (Intelligence Advanced Research Projects Activity).[3] The full dataset 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 was whether Serbia would be granted European Union candidacy in the next four months. The question resolved as "no" because the event did not occur by that date.

As long as a question was open, forecasters were encouraged to update predictions in response to new information. If forecasters did not update on a given day, we assumed 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 questions were open for varying periods and not all forecasters predicted all events,

[2]The aggregate is based on predictions that can become linearly dependent, leading to a degenerate distribution. To illustrate, suppose we model 10 forecasters' predictions jointly with their equally weighted average aggregate. Given that the aggregate is an exact linear combination of the predictions, the individual predictions and the aggregator span only 10 dimensions – not 11. As a result, the joint distribution is degenerate, which causes estimation problems. A common solution is to break this exact linear relationship by adding a small amount of mean-zero noise to the aggregate prediction. In machine learning, adding noise to training data often improves convergence and reduces generalization error (Bishop, 1995). For instance, Cross et al. (2018) "blur away" the discreteness of the data by adding a small amount of white noise to each prediction. In this article, we add normally distributed noise with mean 0 and standard deviation 0.2 to each individual prediction and aggregator in the probit space. This increases the noise level of the crowd and the aggregator by a small amount. However, because this increase occurs for both the crowd and aggregator, it cancels out when we compare the aggregator to the crowd and compute the difference between the noise levels. Adding noise thus allows the estimation procedure to converge, but it does not affect the results noticeably.

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

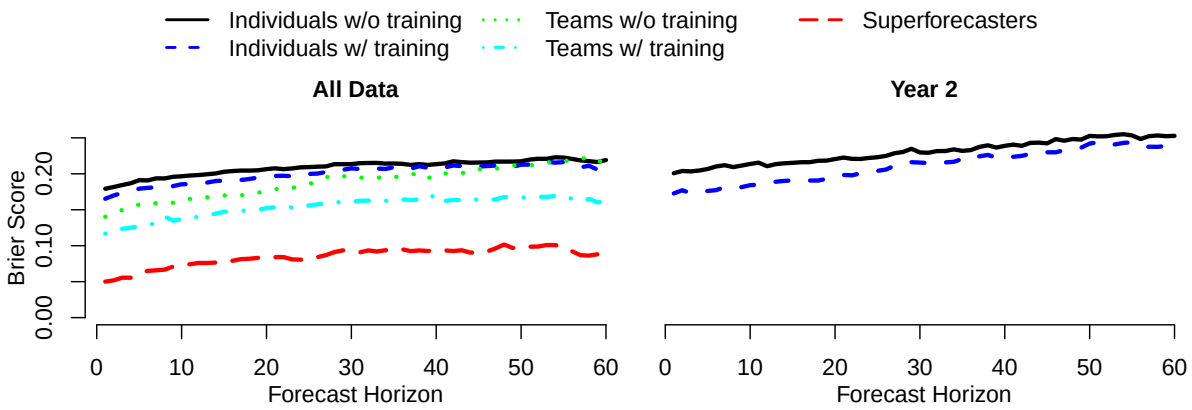

Figure 2: Average individual-level Brier scores in experimental conditions across forecast time horizons.

the number of questions and forecasters per question differ across forecast time horizons and experimental conditions. For more details, see section S1 in the Supplementary Material.

Our analysis focuses on questions with binary outcomes (yes/no) that were open no more than 180 days. This gives us a large set of comparable questions. To ensure forecaster bias has a consistent interpretation, we adjusted questions so that "yes" always refers to a change from the status quo. Forecasters are thus predicting probabilities of change, and bias is either systematic over- or under-estimation of change.

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

- Individuals with and without training in probabilistic reasoning. In the training condition, forecasters completed a tutorial on probabilistic reasoning, drawing on recommendations from the forecast-elicitation literature (O'Hagan et al., 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 prediction.

- Superforecasters. Forecasters' performances were tracked over time. At the end of each tournament year, the top 2% forecasters were designated "superforecasters" and given an opportunity to work together the following year (Mellers et al., 2015). Again, each forecaster ultimately made their own predictions.

Figure 2 shows individual forecasters' average Brier scores in different experimental conditions and at different forecast time horizons defined as the number of days between the prediction and the event resolution. The left panel considers all data, whereas the right panel focuses only on data from year 2. We make this distinction because we can analyze prediction markets only on data from year 2, where the individuals with and without training form appropriate control groups. All other aggregators, however, can be analyzed across years. Forecasters in different experimental conditions exhibit significant heterogeneity in skill. Both

training and teaming improve individual accuracy. Among the five conditions, superforecasters are the most accurate by a substantial margin.

Our goal is to investigate how and to what extent aggregators improve these average individual-level Brier scores. Our experimental conditions allow us to assess the relative value-added of various aggregators for untrained and trained individuals, untrained and trained teams, and superforecasters. Given that teams engage in both informational influence (constructive conversations edging closer to the truth, which ramps up the signal) and normative influence (conformity cascades, which ramp up bias and noise), we expect mixed benefits for teaming for all but the most skillfully run teams (which tend to consist of superforecasters). In the next 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 organize aggregators into two categories: 1. Unsupervised one-off aggregators that can be applied to an isolated prediction task without past performance data; 2. Supervised aggregators that can be applied only if historical performance data are available.

Unsupervised aggregators: The unsupervised aggregators examined include:

1. Probability average (Stone, 1961). In a theory paper, Satopää (2021a) explains that measures of central tendency treat forecasters' disagreements 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 poses a problem because a simple average of probabilities is often too close to 0.5 (Baron et al., 2014). Probabilities are bounded within the unit interval, which means that noise will push large probabilities close to 1.0 downward and push small probabilities close to 0.0 upward.
2. Trimmed probability average (Jose et al., 2014). Trimmed average removes a user-specified percentage of the lowest and highest the probabilities before averaging the remaining probabilities. An alternative to trimming is Winsorizing that 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 too close to 0.5.
3. Median of the probabilities (Hora et al., 2013). This aggregator uses extreme trimming because the median excludes all but the middle value of the individual forecasts.
4. Probit average (Satopää et al., 2014). To debias the simple probability average, we transform the probabilities to their probits in the unbounded real line, average the probits, and then transform the average probit back to the probability scale. This, however, is still a measure of central tendency. Therefore, like the probability average, trimmed probability average, or the 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 2021b). 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 how noisy their judgments are. It is based on a constrained version of the BIN model. Unlike the BIN model, this does not incorporate bias or allow the forecasters possess mutually correlated irrelevant signals. These choices are motivated by the one-off context where each forecaster provides only a single prediction, and 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 separate disagreement due to noise and information asymmetry perfectly based on one prediction per forecaster, extreme levels of disagreement are only possible in the presence of noise. This allows RBA to estimate a lower bound on the level of noise and reduce the risk of overfitting.

   Satopää (2021b) concludes by a model comparison that uses the GJP data to benchmark RBA against other unsupervised aggregators, including the ones mentioned above. This is a horse-race comparison, designed to illustrate the potential of RBA and does not involve any empirical attempts to estimate the BIN pathways via which the different aggregators improve accuracy.

   In addition to the predictions, RBA inputs the forecasters' common prior that 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 Good Judgment Project did not routinely report their prior beliefs. To include RBA in our comparison, we apply it with a default non-informative common prior 0.5.[4] 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. Prediction markets (PM; 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 is implying the event has a 0.6 probability. If forecasters believe 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, causing appropriate adjustments to the current

[4]Satopää (2021b) explains that at the beginning of a question, when uncertainty is the 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. The author then uses the simple average of all probability predictions on the third day of a question as a "compromise" common prior. In section S2 of the Supplementary Material, we show that this leads to very similar results compared to the use of the uniform prior 0.5.

price.

In classic micro-economic theory, prediction markets should be superior to averaging techniques (that only reduce noise) and RBA (that reduces noise and merges partial information). Indeed, in theory, markets should be the most efficient possible method for improving all BIN components. And given that this method relies not on statistical aggregation but on crowds aggregating their own predictions, it sheds light on the pros/cons of statistical aggregation.

The six unsupervised aggregators above have been ordered roughly by the number of BIN components (bias, information, and noise) that, in theory, they should improve. However, this comes at a cost in 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, prediction markets require a trading platform and liquidity management.

Supervised aggregators: Supervised aggregators depend on unknown variables that must be inferred from historical performance data. Which forecaster was closer to right about what? Here we consider a range of supervised aggregators and evaluate them by performing leave-one-out cross-validation. First, we form a testing set by separating one outcome and its predictions from the rest of the data. Next, we estimate the unknown variables of the supervised aggregators using the rest of the data (the training set). Finally, with our estimated model, we predict the event in the testing set by aggregating the predictions in the testing set. Each aggregate prediction is then out-of-sample and has been trained on equal but maximal amounts of training data, so our results are more likely to represent inherent characteristics of the aggregation components.

We break the supervised aggregators into "modules" that address different facets of aggregation. Building on Atanasov et al. (2017), we focus on the effects of three modules that are controlled by 5 unknown variables, denoted ν_1, ν_2, ν_3, ν_4, and ν_5:

1. Temporal decay (TD): We calculate the probit average only based 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 $\nu_{1,min}$ is set so that each aggregate involves at least 3 predictions.

   Our hypothesis is that temporal decay will increase the information levels of the aggregate because forecasters accumulate information over time. Excluding older predictions thus avoids diluting the aggregate with less informed predictions and, instead, gives more weight to the most current, presumptively informed, predictions.

2. Differential weighting (DW): Suppose there are K forecasters, so the equally weighted average would assign 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 on the question. A forecaster k receives 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[5] to predict the outcomes in our training set; $w_{k,upd} \in [0,1]$ is proportional[6] 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 accuracy and frequency in forecast updating, respectively.

This approach has the potential to produce improvement through all three pathways. First, it gives high weight to forecasters who have done well on past prediction tasks. Such performance, of course, suggests these forecasters have low levels of bias and noise, or larger amounts of information. Second, greater updating requires effort, and those who put in extra effort searching for relevant signals in the news, correcting biases and reducing noise are likely to be better forecasters (Atanasov et al., 2017).

3. Recalibration (RC): We transform the forecasters' average probit prediction with a linear function. Specifically, 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 Figure 1) this mechanism can both shift and rotate the points and polarize the distribution of predictions, allowing recalibration to improve all three BIN dimensions.

Note that the parameters of this approach can revert to the probit average if the training data support it. In addition to nesting the probit average, this approach treats the model as modular: Each of the three modules above are includable or removable. For instance, consider a model that only recalibrates the probit average—and does not perform differential weighting or temporal decay. The only module in this model would be recalibration, which we denote as RC. Alternatively, we could recalibrate the probit average with differential weights. This model would be called RC + DW. Mixing and matching modules in this manner leaves us with 7 distinct combinations of modules (TD, DW, RC, TD+DW, TD+RC, DR+RC, and TD+DW+RC) and hence 7 distinct aggregators. By training each of these aggregators, we can study the marginal effect of each module.

[5] We measure accuracy as $1 - \mathrm{BrS}_k$, 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 happened, then $1 - \mathrm{BrS}_k = 1 - \frac{1}{2}[(1-0.5)^2 + (1-0.7)^2] = 1 - 0.17 = 0.83$. 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 way, forecasters' past predictions will determine whether a forecaster is treated as net helpful (accuracy higher than 0.75) or net harmful (accuracy lower than 0.75).

[6] We count how often forecasters update their predictions for the test event and then normalize all counts by the maximum number of updates by each forecaster.

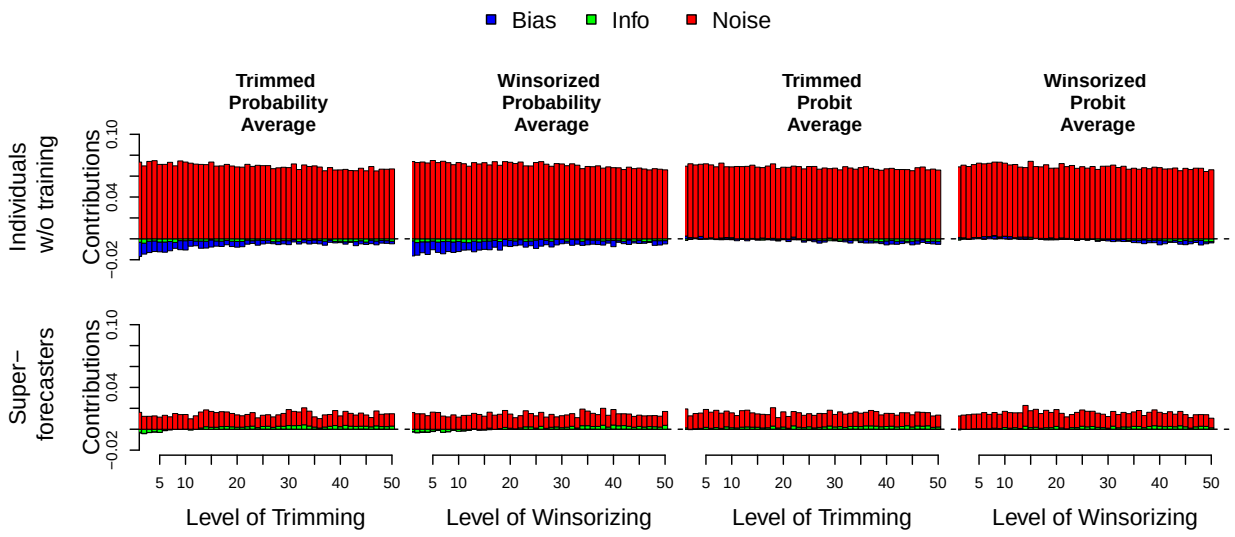

Figure 3: Brier score contributions at different levels of (symmetric) trimming and Winsorizing of probability and probit predictions.

6. Results

6.1. Unsupervised: Statistical Aggregators

Our first analysis involves unsupervised aggregators, except for prediction markets to which we will return later for two reasons. First, all unsupervised statistical aggregators, such as probability average and median, can be applied to forecasts across all four years of the tournament. So, we will maximize power with the statistical aggregators using data from all four years. Then, we will compare prediction markets 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 can study their behavior in a diverse set of environments. Unfortunately, the same is not true for prediction markets. For instance, based on past data, we cannot study prediction markets of different sizes. The market is a fixed entity that cannot be broken into pieces in this fashion. But we can apply the statistical aggregators to randomly sampled crowds of different sizes. In addition to crowd size, we will inspect 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) show that trimming or Winsorizing forecasters' probabilities before averaging can boost accuracy. To understand how, we apply different levels of trimming and Winsorizing to probability and probit predictions made 30 days before event resolution and average the remaining predictions.

Figure 3 shows improvements in the individual-level expected Brier scores (recall Figure 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 of individual forecasters, as indicated by labels on the left. The y-axis of each subplot measures change in expected Brier score. Vertical colored bars show how much the aggregator delivers improvement via bias, information, and noise. The x-axis of each subplot represents the level of trimming or Winsorizing, which grows from left to right.

The left-most column in Figure 3 represents the trimmed probability average. The top figure in this column shows that averaging probabilities with no trimming improves untrained individuals' expected Brier scores by 0.07 through noise reduction. But the probability average also increases bias and reduces information. The combined effect of the increased bias and decreased level of information is a 0.02 increase in the expected Brier score. Therefore, the net improvement in expected Brier score is about 0.07–0.02 = 0.05. The negative effect of bias, however, can be almost eliminated by enough trimming, as one sees by moving right along the x-axis.

The effect of Winsorizing probabilities before averaging is similar to the effect of trimming, as shown in the second column from the left. The final two columns consider the effects of trimming and Winsorizing probits before averaging them. As before, these mechanisms have very similar effects. Contrary to the probability average, however, too much trimming or Winsorizing before averaging the untrained individuals' probits slightly increases the level of bias. The bottom row shows that trimming or Winsorizing has a negligible effect on the probability average and the probit average of superforecasters' predictions. This suggests that trimming and Winsorizing are less effective for highly skilled forecasters.

In practice, however, there is no principled way of 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 have very similar effects on the probability average, for the rest of this article, we will only consider the trimmed probability average with 10% trimming, which we see as a likely default choice.

6.1.2. Number of Forecasters

Figure 4 resembles Figure 3 except that, instead of analyzing the effects of trimming or Winsorizing, the focus is on unsupervised statistical aggregators and how they improve individuals' accuracy (recall Figure 2) based on predictions made by crowds of varying size. The x-axis of each subplot represents crowd size, which grows from 5 on the left to 75 on the right. Again, we fix the forecast time horizon to 30 days before event resolution. To capture trends for each size and aggregator, we form 250 different crowds by randomly sampling forecasters, calculate the accuracy improvement due to aggregation under each crowd, and plot the average BIN decomposition.

The results show that averaging techniques (first four columns from the left) can benefit from additional predictions if the crowd consists of untrained individuals. This improvement, however, is rather small and comes largely from better noise reduction. Unfortunately, the larger crowd size does not help averaging to overcome its bias toward 0.5.

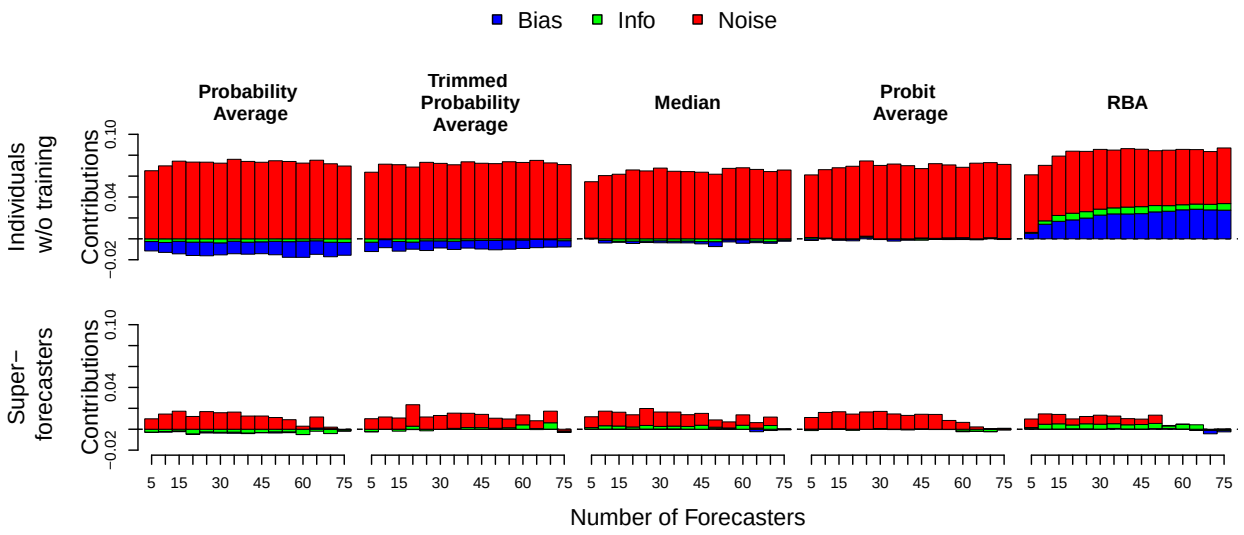

Figure 4: Average Brier score contributions from unsupervised statistical aggregators under varying sizes of the crowd of forecasters.

By contrast, RBA draws bigger benefits from larger crowds of non-trained individuals, as shown by the right-most column. Satopää (2021b) explains that RBA assumes larger crowds to possess more total information. Therefore, as the crowd size increases, RBA seeks to incorporate increasing amounts of information and ends up extremizing the consensus prediction, given by the probit average, more heavily away from the forecasters' common prior (in our case 0.5). Our results show that the resulting aggregate is more informed and less biased than the typical individual forecaster. It is perhaps surprising that RBA can reduce bias even though its model does not acknowledge bias in individual predictions. But an aggregator does not need to model bias to reduce it. If individual forecasters are biased, an increasingly accurate aggregator must eventually reduce this bias. We return to this topic in subsection 6.1.4. A large number of predictions, however, is not necessary for RBA to boost the individual forecasters' accuracy. Indeed, even with just 5 predictions, it improves accuracy as much as the best averaging technique, the probit average. And it does this by tapping into all three BIN dimensions.

6.1.3. Level of Information Sharing

Next, we consider crowds who differ in the level of information shared among individual members. We sample individuals who work alone (with or without training) and place them in the same crowd with individuals who were known to work on the same team (in the teaming or superforecaster condition, respectively). A 10-person crowd of, say, 3 untrained forecasters working alone and 7 untrained forecasters on the same team represents more information sharing than a crowd of 7 untrained forecasters working alone and 3 untrained forecasters on the same team.

Our analysis fixes the forecast time horizon at 30 days, considers crowds of size 10, and varies the number of non-team members from 0 to 10. To capture general trends, for each number of non-team members and

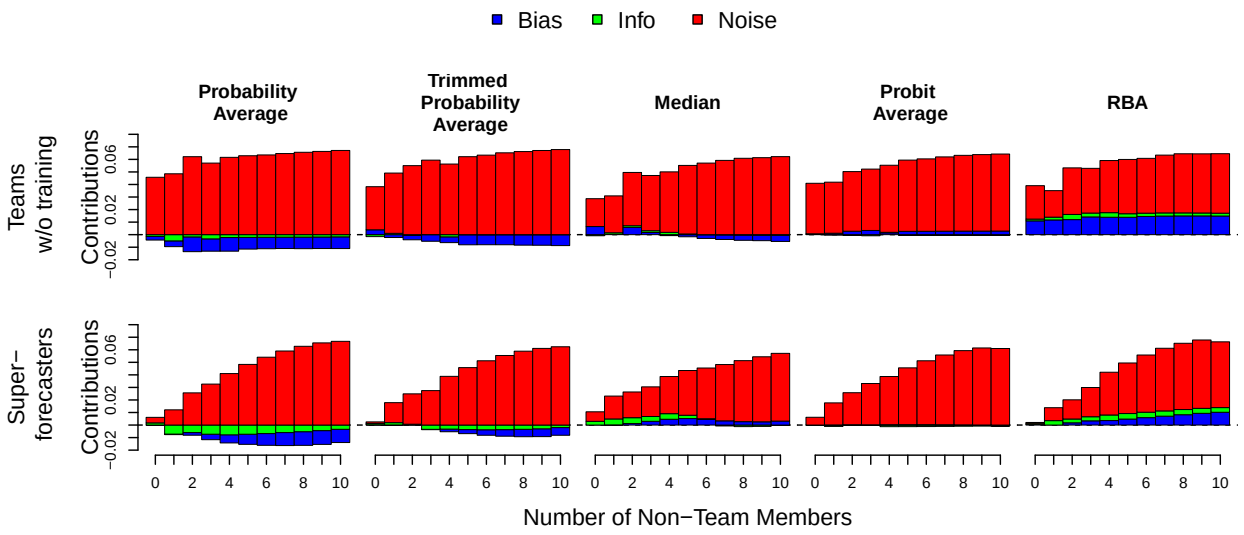

Figure 5: Average Brier score contributions from unsupervised statistical aggregators under different levels of information sharing among crowd members.

aggregators, we form 250 different crowds by augmenting randomly chosen teams with individual forecasters working alone, calculate the accuracy improvement due to aggregation under each crowd, and plot the average BIN decomposition in Figure 5. Here the x-axis of each subplot represents the number of non-team members, which grows from left to right. Therefore, crowds with more information sharing among members are on the left side of each plot. To maintain a similar dataset at each number of non-team members, we only consider questions that involve both non-trained individuals and teams. For the number of teams and questions available at each number of non-team members, see section S1 in the Supplementary Material.

The results show that the less the crowd members interact with each other, the better RBA and probit average perform relative to the other aggregators that lose accuracy due to increasing bias. Forecasters that work alone are likely to make more different predictions, which can cause the average to be too close to 0.5.

As the crowd members share more information, noise reduction becomes less important. This happens because forecasters in the teaming condition have less noise (Satopää et al., 2021), leaving less room for noise reduction. This effect is particularly strong for RBA that performs worse than probit average when the crowd consists almost entirely of team members. In general, RBA seeks to model disagreement due to noise and information asymmetry in the crowd. In a one-off setting with limited data, however, 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, leading to overfitting, which adds noise. RBA, however, also performs noise reduction. The net effect of noise reduction and overfitting is still positive, leading overall to less noise and thus more accuracy. However, due to overfitting, RBA is less effective at reducing noise than simpler

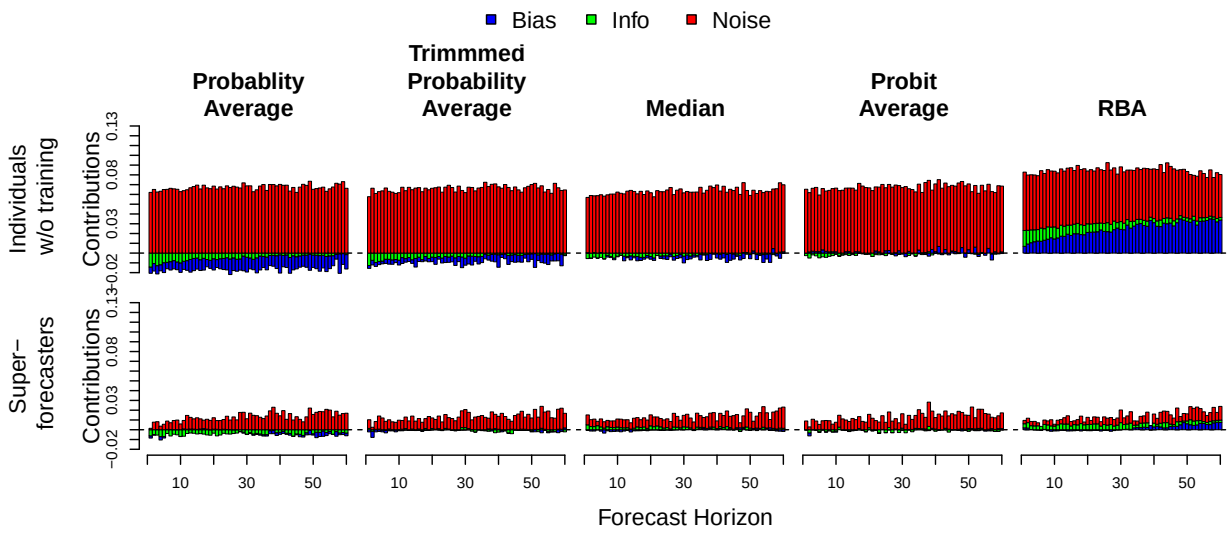

Figure 6: Brier score contributions from unsupervised statistical aggregators under different forecast time horizons.

averaging techniques that assume all disagreement stems from noise – an assumption that aligns better with a crowd of team members. We will return to overfitting in our discussion of supervised aggregators.

6.1.4. Forecast Time Horizon

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

Focus first on the top row, showing how aggregators improve the accuracy of individuals without probabilistic training. Averaging techniques do this almost entirely via noise reduction. The simple probability average increases bias of longer-range predictions, but this is largely 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, a form of extreme trimming, can alleviate these downsides. Probit averaging appears to correct this undesirable tendency almost entirely.

All these aggregators, however, are measures of central tendency and hence treat forecasters' disagreement as noise. So, whenever there is information asymmetry, these aggregators may, in principle, cause information loss. This effect is visible at short forecast time horizons (moving left on each subplot) and shrinks as we move column by column from left to right, so that the probit average reduces information very little if at all. 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, like the probability average, median, or trimmed probability average.

Overall, the regularized Bayesian aggregator (RBA) is the most successful at improving accuracy, as shown by the height of the colored bars. Furthermore, unlike measures of central tendency that act almost

exclusively through noise reduction, RBA delivers improvements across all three BIN-dimensions: bias, noise, and information. About half of the improvement comes from noise reduction. For long forecast time horizons (toward right on each subplot), RBA's accuracy improves through bias reduction more than through merging individual forecasters' dispersed information. This ordering, however, reverses as the horizon shortens (toward the left side of the plots).

To understand this trend, 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. It is when event resolution nears that 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 case, consider the probit average. This affects accuracy entirely through noise reduction and preserves the individual forecasters' level of information and bias. In contrast to the probability average that is often too close to 0.5, RBA seeks to integrate forecasters' dispersed information and typically[7] extremizes the probit average away from the user-specified common prior that in our case is set at the non-informative prediction 0.5. Our results now show that such extremization reduces the forecasters' bias, particularly for long forecast time horizons. Therefore, even though the probability average is systematically too close to 0.5, so are the individual forecasts and extremizing their predictions directly away from 0.5 can reduce this bias.

Finally, consider the bottom row, showing how aggregators improve the superforecasters' accuracy. This shows that averaging performs better when individuals are highly skilled. For instance, superforecasters may be approaching the limits of epistemic uncertainty so there is not much public information to incorporate into their forecasts and remaining disagreements are likely to stem from noise, which can be reduced by averaging. And our results suggest that the best one can do is to average superforecasters' predictions. The differences among aggregators here are small and not statistically significant, as is 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 prediction markets run by Lumenogic. Participants worked independently. In one market, they received training in how

[7] 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. In our current analysis, however, the crowd size is typically larger than 50 (see the Supplementary Material). For such 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 Satopää (2021b).

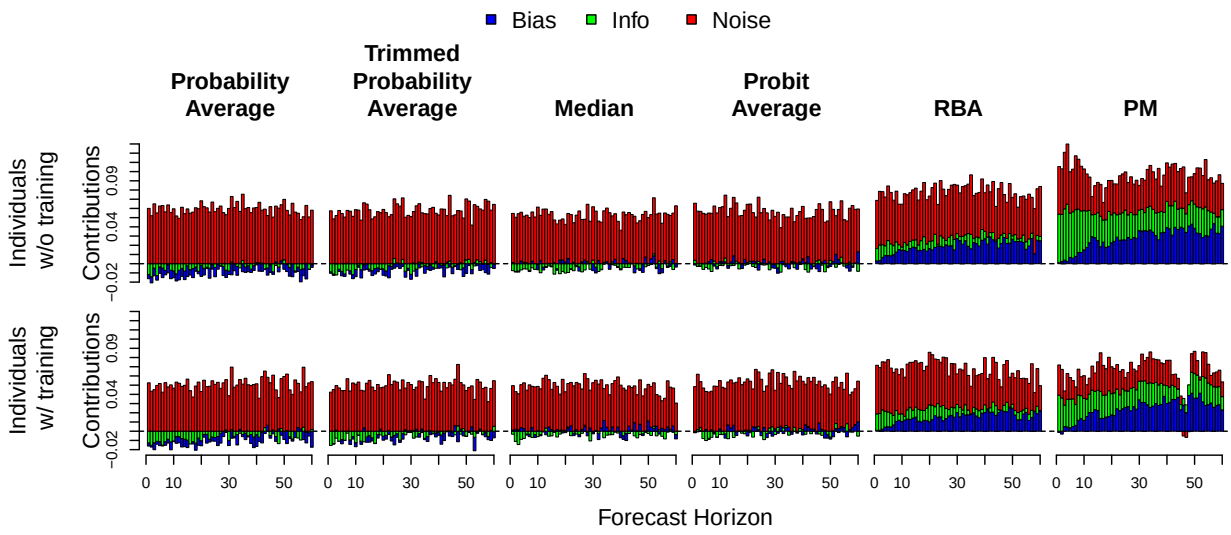

Figure 7: Brier score contributions from unsupervised aggregators, including prediction markets, under different forecast time horizons.

to maximize profits and avoid market errors. In the other market, there was no training.[8] Therefore, in this subsection, we only consider individuals with and without training as the control experimental conditions.

As before, Figure 7 shows improvements in the individual-level expected Brier scores due to aggregation at different forecast time horizons and traces that improvement to bias reduction, noise reduction, and information acquisition. Given that the analysis uses much less data than in Figures 3-6, there is much more variability between consecutive colored bars, leading to more rugged plots.

For the unsupervised statistical aggregators, however, the results agree qualitatively with those in the previous section: averaging improves accuracy almost entirely through noise reduction whereas RBA improves through all three BIN-dimensions. The novel component is the Prediction Market (PM) column on the right which reveals that PM improves accuracy via all BIN dimensions. Furthermore, it improves bias and information levels noticeably more than RBA, leading to the highest reduction in individual-level expected Brier score among all unsupervised aggregators in this article.

6.3. Supervised: Statistical Aggregators

6.3.1. Forecast Time Horizon

Our next analysis involves supervised aggregators that require past performance and outcome data. In practice, statistical models are rarely built at once, and the best models are those tailored to the problem at

[8]The third year had several predictions markets, also run by Lumenogic. Instead of separating trained and untrained individuals into separate markets, all individuals participated in the same market. This makes it impossible to separate the effects of training from other factors. Therefore, to simplify the discussion, we focus only on data from the second year of the tournament when we had a clear control group for each prediction market.

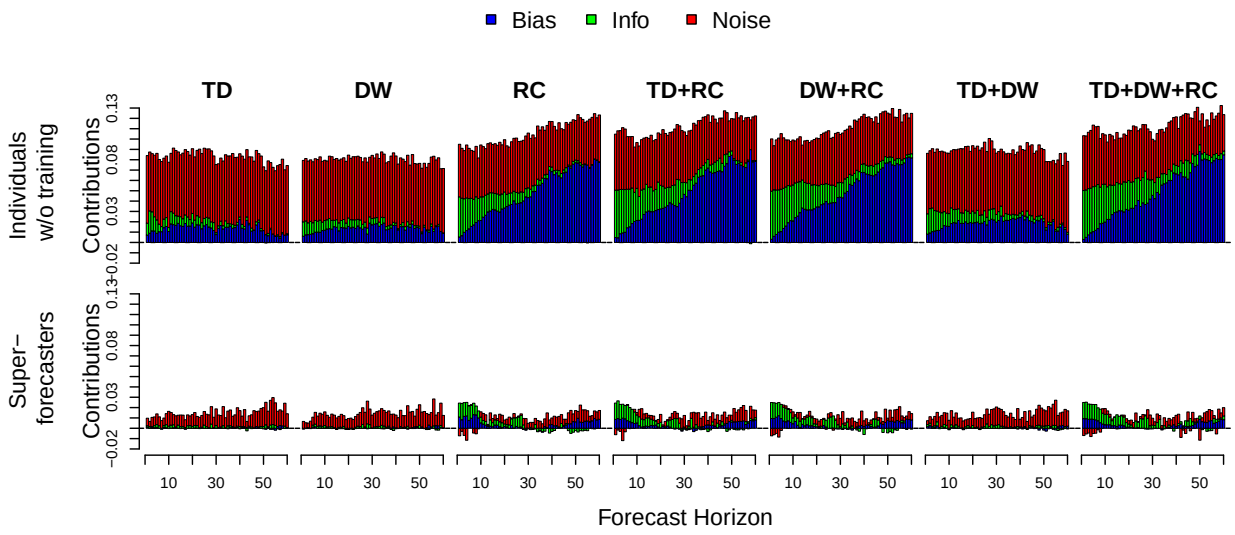

Figure 8: Brier score contributions from supervised aggregators with different combinations of the modules under different forecast time horizons.

hand. It often makes sense to begin with a simple model, spot problems, and look for incremental improvements (Smith, 1984; Draper, 1995; Schad et al., 2020). To this end, we propose a modular construction of aggregation in which one tries different modules and finds 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, consider building a racing car. The top speed of your model is on par with other teams' models, but poor acceleration is keeping you off the podium. The braking system is unlikely to improve acceleration, so your team decides not to prioritize brakes. Rather, it priorities the engine. Our modular construction of the aggregator allows a similar understanding of the modules. For instance, one may find that differential weighting improves accuracy through noise reduction. However, there are many ways to perform such weighting. With our analysis, one can then make targeted changes and look for the desired effect in a single module. One may find that the aggregator improves the forecasters' overall accuracy, but one module is limiting that potential and, by removing that module, the aggregator performs even better. One can also inspect interactions between modules. One may find that some modules work well or poorly together. The bottom-line is that the modular construction makes development of aggregators more systematic and robust.

To illustrate this idea, we consider all possible mixes of the three modules described in section 5. As Figures 6-7 did for unsupervised aggregators, Figure 8 does for supervised aggregators: it decomposes the accuracy improvements into bias reduction, noise reduction and enhanced extraction of forecasters' dispersed information at different forecast time horizons. The seven columns represent all possible mixes of the three modules.

First, consider the three left-most columns in Figure 8, corresponding to models with individual modules

(TD, DW, and RC). All three modules improve accuracy through noise reduction. Each one builds on the probit average that is an effective way to reduce noise in individual predictions. A comparison of the probit average in Figure 6 to temporal decay (TD) and differential weighting (DW) in Figure 8 shows that these modifications to the probit average can reduce bias and combine forecasters' dispersed information. Given that both TD and DW are weighted averages, they can achieve this by assigning more weight to forecasters who are expected to have lower bias and more information. The only visible change from TD and DW to TD+DW is the slightly increased level of information in TD+DW. This suggests that TD and DW capture somewhat different parts of forecasters' dispersed information.

Measures of central tendency, like the weighted average driving TD and DW, however, are limited in the extent to which they can shift predictions systematically up or down, which could correct a bias, or make predictions more extreme, which could increase the level of information (recall Figure 1). By contrast, RC is less limited and can deliver stronger bias reduction and information fusion than TD and DW. Aside the slightly increased level of information in the more complex models, this 1-module RC model performs as well as or better than the more complex 2- or 3-modules models (in the 4 right-most columns).

As Albert Einstein once said: "Everything should be made as simple as possible, but not simpler." Based on this principle, essentially a refinement of Occam's razor, we should look for the simplest model without compromising too much accuracy. Even though TD and DW capture slightly different shares of forecasters' information, the marginal benefit in accuracy from including both modules may not be worth the added computational burden to estimate their parameters. In this sense, including both TD and DW leads to redundancy. A decision maker may then prefer the simpler 3-parameter TD+RC over the 4-parameter DW+RC or 5-parameter TD+DW+RC that result in similar degrees of accuracy.

This analysis is illuminating but it does not shed light on how the modules interact and contribute to accuracy improvements. To understand this, we can generalize the BIN decomposition in Satopää et al. (2021) by further splitting the three BIN modules into the effects brought by RC, TD, and DW. This illustrated in Figure 9 that turns up the microscope on the right-most column of Figure 8 (TD+RC+DW) and shows how each module contributes to the accuracy of this 3-modules model. From left to right, the three columns now correspond to bias, information, and noise contributions. As before, rows are different experimental conditions of individual forecasters, as labeled on the left. 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 left-most column shows that essentially all bias reduction comes from RC and that this contribution decreases along the forecast time horizon. Similarly, the middle column shows that in the 3-modules model most of the information acquisition occurs via RC. And the right-most column shows that the task of noise reduction in the aggregator is split almost equally between TD and DW.

Interestingly, recalibration (RC) may introduce additional noise at long forecast time horizons (toward the right side of each plot in the right-most column). In section S4.1.2 of the Supplementary Material, we show that this increase in noise is particularly pronounced when we aggregate predictions from untrained

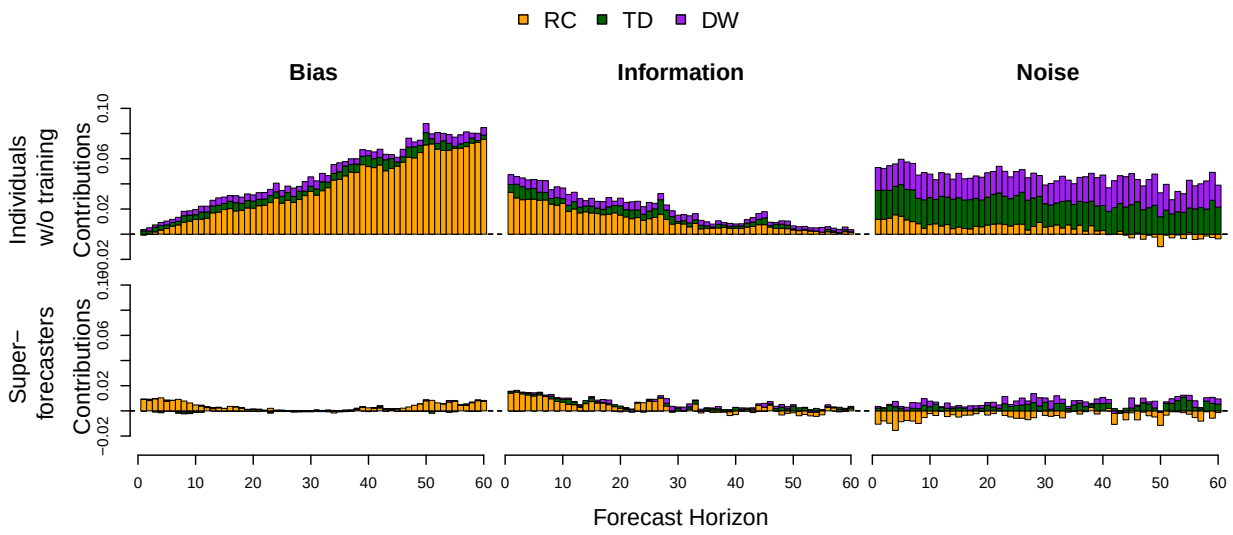

Figure 9: Detailed decomposition of the improvements due to the different modules in the full model TD+DW+RC from bias reduction, information fusion, and noise reduction.

teams, perhaps because this condition is where we have the least data. One explanation for this pattern is overfitting: The recalibration mechanism falsely interprets irrelevant signals as information. This illustrates how merging forecasters' information can be a double-edged sword. Perfect fusion requires the aggregator to distinguish two types of variability: that which covaries with outcomes (information) and that which does not (noise). Any mistake leads to overfitting. Averaging, by contrast, is less ambitious. It does not seek to merge dispersed information. Instead, it takes the conservative approach of treating all variability as noise. This avoids overfitting but at the price of excessive conservativism. Winkler et al. (2019) make a similar distinction. They explain that sophisticated techniques can produce more accurate predictions but are prone to overfitting. In this sense, they are riskier than simple rules such as averaging.

Returning to our racing-car analogy, Figure 9 suggests a way of improving the 3-module model: one could make the recalibration module robust against over-fitting by regularizing the estimation of the recalibration parameters ν_4 and ν_5 with elastic net or other related techniques (Zou and Hastie, 2005). Implementing such approaches, however, is more complicated and outside the scope of the current article. Note, however, that this direction for improvement is only visible in Figure 9 and could not have been deduced from the less granular analysis in Figure 8.

7. General Discussion

The current study demonstrates the value of moving beyond horse-race comparisons centered on which aggregators reduce error scores the most and instead isolating the pathways by which promising methods improve accuracy (Soll and Larrick, 2009; Davis-Stober et al., 2014). Our empirical findings largely matched our theoretical expectations, with the exceptions noted below.

As expected, averaging techniques (unweighted or weighted) boosted accuracy almost entirely by noise reduction. Simple averaging also increased bias largely 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 principled way to choose the level of trimming in a ex-ante one-off forecasting context. Our results, however, suggest that the probit average offers a simple alternative that can correct this tendency without the need to choose values of tuning parameters. That said, all of these aggregators produce measures of central tendency and may, at least in theory, reduce information because they treat all disagreement as noise, thus eliminating disagreements that arise from different sets of information. Our empirical analysis encountered this pattern for all averaging techniques, except the probit average, which—broadly speaking— preserves forecasters' level of bias and information and improves accuracy purely through noise reduction.

Better bias reduction and fusion of information then offer opportunities for further improvement. But capturing these opportunities will not come free: it will entail increased complexity and data requirements. In particular, Dietrich (2010) explain that combining forecasters' dispersed information with a statistical aggregator is not possible without their prior beliefs. On the methodology side, Satopää (2021b) devised a Regularized Bayesian Aggregator (RBA) that is estimated with numerical Markov chain Monte Carlo techniques and inputs forecasters' common prior and predictions. Although the Good Judgment Project did not collect forecasters' prior beliefs, we illustrated how using a non-informative uniform prior 0.5 still allowed RBA to improve accuracy by tapping into all three BIN dimensions. We suspect that further improvements could be achieved by using forecasters' actual prior beliefs.

All of this suggests that, by introducing more rigor and discipline, the state-of-the-art practice of crowd wisdom can move well beyond mere averaging and noise reduction. As long as we continue eliciting forecasters' predictions alone, statistical aggregation is focused on noise reduction. It is time to consider a more principled process where forecasters first seek an appropriate base rate and then update it to take into account their context-specific information. This entails more effort but can reduce bias by centering the individual predictions more appropriately and then allow aggregators, such as RBA, to reduce noise and merge the forecasters' dispersed information.

RBA, however, did not increase information extraction as much as the prediction markets. Prediction markets are a lot more complicated and often expensive to implement than averaging or RBA but also deliver a broader range of benefits. As expected, they boosted accuracy via all three components of the BIN model. Prediction markets are often well calibrated (Atanasov et al., 2017), which suggests that they reduce bias and noise. Among the aggregators we considered, prediction markets proved best at improving the extraction of (partial) information. This result is consistent with micro-economic arguments about the power of markets to facilitate effective information sharing among individuals. One possibility is that, at least in this context, information aggregation is best handled by market participants, not mechanical aggregators. However, statistical aggregation is slightly more efficient in reducing bias and noise. Humans working in

markets may be better at extracting information in geopolitical tournaments and machines better at tamping down noise and bias (Atanasov et al., 2017).

This suggests a strategy that allows decision makers to combine the forecasters' dispersed information without using complex aggregators such as RBA or prediction markets. The idea is to alter the forecasters' level of bias, information asymmetry, and noise to better match averaging. For instance, allowing forecasters to exchange information in teams makes their predictions more amenable to averaging. In a case study of an anonymized electronics company, Oliva and Watson (2009) explain 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 base their predictions on this common information. This information management strategy reduces information asymmetry and should boost the performance of averaging. Alternatively, the decision maker can carry out a noise audit (Kahneman et al., 2021) that directs forecasters to decompose the signals they are using in making probability judgments, independently assess their values and then weight 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 to another. Flushing these disagreements into the open can reveal mistakes and after adjusting predictions, forecasters might be able to reconcile their disagreements, driving down 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, the less epistemic uncertainty there is, the less important information fusion is, and, conversely, the more appropriate averaging is. Superforecasters in the rather noisy domain of geopolitics might even be approaching the limits of epistemic uncertainty. At first, it seems plausible that a superforecaster may have acted strategically and chosen to not share some private information with the others. However, according to our results, the best one can do is to average the superforecasters' predictions. Except at short forecast time horizons, averaging superforecasters' predictions performed as well as even the supervised techniques that have access to past performance and outcomes data. Therefore, most if not all of the superforecasters' disagreement is likely to stem from noise, suggesting that very little information was actually held back and that there may not be much more useful public information to incorporate 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 from information fusion because there was less information asymmetry among forecasters. Few possessed insider information about how, say, the Syrian civil war would end when it started in 2011 or how a U.S. Presidential election would end before the primaries had begun. Forecasters are often roughly equally ignorant about far-off futures. So, most of the gains come from bias or noise reduction where statistical aggregators are most helpful. This discovery suggests a strategy for making long-range predictions: reduce bias by offering the forecasters relevant base rates and then reduce noise by averaging their (probit) predictions.

If the use of base rates can successfully remove bias and results in longer-range predictions that disagree largely due to noise, then averaging can outperform RBA that always assumes some degree of information diversity and hence can overfit. For shorter-range predictions, when relevant information is more available, information asymmetry emerges. Following our discussion above, here the decision maker may look for ways to give the forecasters access to a common set of information before asking for and averaging their predictions. However, if this is not possible and the decision maker is left with predictions that disagree due to information asymmetry, information-fusion techniques, such as RBA, are likely to offer better results than averaging.

This raises the question: at what forecast time horizon should one switch from averaging to something like RBA? The right timing depends on when relevant news begin to emerge and hence is likely to depend on the context. Machine learning techniques can help to track the evolution of such news (e.g., Allan et al. 1998). Alternatively, in the ex-ante context without outcomes data, it may be possible to use Natural Language Processing techniques developed by computational linguists and analyze the forecasters' written rationales behind their predictions (Karvetski et al., 2021). Implementing the mechanics for inferring the forecasters' level of information asymmetry and noise from their written rationales is a challenge that we reserve for future work.

These examples illustrate the complementarities of aggregation and elicitation. However, the benefits extend beyond matching aggregators to environments. Knowing how an aggregator boosts accuracy enhances our understanding of its mechanism and hence can offer guidance towards further methodological improvements. Based on the BIN decomposition, we proposed a modular construction of supervised aggregators which offers a granular view of each module of the aggregator and suggests improvements by pointing out potential redundancies or shortcomings. One could apply a similar approach to an experimental intervention to improve individual forecasters. For instance, consider designing a training program or elicitation scheme that consists of calibration training, outcome feedback, and team discussion. Due to participant fatigue, it may be impractical to incorporate all modules. Some choices must be made. By eliciting predictions under each combination of modules, our method suggests ways to mix-and-match modules and maximize improvement across three dimensions, within the constraints of the participants. Of course, with many modules, the total number of combinations becomes huge, making full factorial designs impractical. But a partial solution is fractional factorial designs that avoid redundancies by carefully choosing subsets of combinations from the full factorial design (see, e.g., Montgomery 2017).

Modular construction of aggregators, however, does not face the same challenge in data collection. Here we can simply collect predictions and apply all possible combinations of aggregation modules to the data, as we did in section 6.3. The analysis can reveal redundancies, prune unnecessary parameters and develop unsupervised aggregators in one-off contexts where data are often limited. In such contexts, the model must find a balance between the important aspects of the problem and what can be estimated from data (Palley and Satopää, 2020). Although it is currently unclear what the modular construction of a one-off aggregator

could be, it is an intriguing challenge for future research.

References

Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y., 1998. Topic detection and tracking pilot study final report .

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

Atanasov, P., Rescober, P., Stone, E., Swift, S.A., Servan-Schreiber, E., Tetlock, P., Ungar, L., Mellers, B., 2017. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. Management science 63, 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, 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, 267–280.

Chen, Y., Pennock, D.M., 2010. Designing markets for prediction. AI Magazine 31, 42–52.

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. International journal of forecasting 5, 559–583.

Clemen, R.T., Winkler, R.L., 1999. Combining probability distributions from experts in risk analysis. Risk analysis 19, 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, 259–268.

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

Dietrich, F., 2010. Bayesian group belief. Social choice and welfare 35, 595–626.

Draper, D., 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society: Series B (Methodological) 57, 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, 519.

Hora, S.C., Fransen, B.R., Hawkins, N., Susel, I., 2013. Median aggregation of distribution functions. Decision Analysis 10, 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, 463–475.

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

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, e0194889.

Mannes, A.E., Soll, J.B., Larrick, R.P., 2014. The wisdom of select crowds. Journal of personality and social psychology 107, 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, e1514.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., et al., 2015. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science 10, 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., Oakley, J.E., Rakow, T., 2006. Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons, Chichester.

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, 138–151.

Palley, A., Satopää, V., 2020. Boosting the wisdom of crowds within a single judgment problem: Selective 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, 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., 2021b. Regularized aggregation of one-off probability predictions. Available at SSRN: https://ssrn.com/abstract=3769945.

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, 344–356.

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

Satopää, V.A., Salikhov, M., Tetlock, P.E., Mellers, B., 2021. Bias, information, noise: The bin model of forecasting. Management Science 0. Available at https://doi.org/10.1287/mnsc.2020.3882.

Schad, D.J., Betancourt, M., Vasishth, S., 2020. Toward a principled bayesian workflow in cognitive science. 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, 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, 780.

Stone, M., 1961. The linear opinion pool. 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, 239–260.

Wolfers, J., Zitzewitz, E., 2004. Prediction markets. Journal of Economic Perspectives 18, 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, 301–320.