Contents lists available at ScienceDirect

International Journal of Forecasting

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

Combining multiple probability predictions using a simple logit model

Ville A. Satopää[a,*], Jonathan Baron[b], Dean P. Foster[a], Barbara A. Mellers[b], Philip E. Tetlock[b], Lyle H. Ungar[c]

[a] Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA, 19104-6340, USA

[b] Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Solomon Lab Bldg., Philadelphia, PA, 19104-6241, USA

[c] Department of Computer and Information Science, University of Pennsylvania, Levine Hall, 3330 Walnut Street, Philadelphia, PA, 19104-6309, USA

ARTICLE INFO

Keywords:
Combining forecasts
Error correction models
Expert forecasts
Logit-normal models
Multinomial events
Probability forecasting

ABSTRACT

This paper begins by presenting a simple model of the way in which experts estimate probabilities. The model is then used to construct a likelihood-based aggregation formula for combining multiple probability forecasts. The resulting aggregator has a simple analytical form that depends on a single, easily-interpretable parameter. This makes it computationally simple, attractive for further development, and robust against overfitting. Based on a large-scale dataset in which over 1300 experts tried to predict 69 geopolitical events, our aggregator is found to be superior to several widely-used aggregation algorithms.

1. Introduction

Experts are often asked to give decision makers subjective probability estimates as to whether certain events will occur or not. Having collected such probability forecasts, the next challenge is to construct an aggregation method that will produce a consensus probability for each event by combining the probability estimates appropriately. If the observed long-run empirical distribution of the events matches that of the aggregate forecasts, the aggregation method is said to be calibrated. This means that, of the events which have been assigned an aggregate forecast of 0.3, for instance, 30% should occur. According to Ranjan (2009), however, calibration is not sufficient for useful decision making. The aggregation method should also maximize *sharpness*, which increases as the aggregate forecasts concentrate more closely around the extreme probabilities 0.0 and 1.0. Therefore, it can be said that the overall goal in probability estimation is to maximize the sharpness, subject to calibration (for more information, see for example Gneiting, Balabdaoui, & Raftery, 2007; Pal, 2009).

The most popular choice for aggregation is *linear opinion pooling*, which assigns each individual forecast a weight which reflects the importance of the expert. However, Ranjan and Gneiting (2010) show that any linear combination of (calibrated) forecasts is uncalibrated and lacks sharpness. Furthermore, in several simulation studies, Allard, Comunian, and Renard (2012) show that linear opinion pooling performs poorly relative to other pooling formulas with a multiplicative instead of an additive structure.

The literature to date has introduced a wide range of methods for aggregating probabilities in a non-linear manner (see for example Bordley, 1982; Polyakova & Journel, 2007; Ranjan & Gneiting, 2010); however, many of these methods involve a large number of parameters, making them computationally complex and susceptible to over-fitting. By contrast, parameter-free approaches, such as the median or the geometric mean of the odds, are too

[*] Corresponding author. Tel.: +1 215 760 7263; fax: +1 215 898 1280.
E-mail addresses: satopaa@wharton.upenn.edu (V.A. Satopää), baron@psych.upenn.edu (J. Baron), dean.foster@gmail.com (D.P. Foster), mellers@wharton.upenn.edu (B.A. Mellers), tetlock@wharton.upenn.edu (P.E. Tetlock), ungar@cis.upenn.edu (L.H. Ungar).

0169-2070/$ – see front matter © 2013 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijforecast.2013.09.009

simple to be able to incorporate the use of training data optimally. In this paper, we propose a novel aggregation approach that is simple enough to avoid over-fitting, straightforward to implement, and yet flexible enough to make use of training data. Thus, our aggregator retains the benefits of parsimony from parameter-free approaches, but without losing the ability to use training data.

The theoretical justification for our aggregator arises from a log-odds statistical model of the data. The log-odds representation is convenient from a modeling perspective. Being defined on the entire real line, the log-odds can be modeled using a Normal distribution. For example, Erev, Wallsten, and Budescu (1994) model log-odds with a Normal distribution centered at the "true log-odds".[1] The variability around the "true log-odds" is assumed to arise from the personal degree of momentary confidence that affects the process of reporting an overt forecast. We extend this approach by adding a *systematic bias* component to the Normal distribution. That is, the Normal distribution is centered at the "true log-odds", which have been multiplied by a small positive constant (strictly between zero and one), and hence, are systematically regressed toward zero.

To illustrate this choice of location, assume that 0.9 is the most informed probability forecast that could be given for a future event with two possible outcomes. A rational forecaster who aims to minimize a reasonable loss function, such as the Brier score,[2] without any previous knowledge of the event, will give an initial probability forecast of 0.5. However, as soon as he gains some knowledge about the event, he will produce an updated forecast that is a compromise between his initial forecast and the new information acquired. The updated forecast will therefore be conservative, and necessarily too close to 0.5, as long as the forecaster remains only partially informed about the event. If most forecasters fall somewhere on this spectrum between ignorance and full information, their average forecast will tend to fall strictly between 0.5 and 0.9 (see Baron, Ungar, Mellers, & Tetlock, submitted for publication, for more details). This discrepancy between the "true probability" and the average forecast is represented in our model by the use of the regressed "true log-odds" as the center of the Normal distribution.

Both Wallsten, Budescu, and Erev (1997) and Zhang and Maloney (2012) recognize the presence of this systematic bias. Wallsten et al. (1997) discuss a model with a bias term that regresses the expected responses towards 0.5. Zhang and Maloney (2012) provide multiple case studies showing evidence of the existence of the bias. However, neither study describes either a way of correcting the bias or a potential aggregation method to accompany the correction. Zhang and Maloney (2012) estimate the bias at an individual level, requiring multiple probability estimates from a single forecaster. Even though our approach can be extended rather trivially in order to correct the bias at any level (individual, group, or collective), in this paper we treat the experts as being indistinguishable, and correct the systematic bias at a collective level by shifting each probability forecast closer to its nearest boundary point. That is, if the probability forecast is less (more) than 0.5, it is moved away from its original point and closer to 0.0 (1.0).

This paper begins with the modeling assumptions that form the basis for the derivation of our aggregator. After describing the aggregator in its simplest form, the paper presents two extensions: the first generalizes the aggregator to events with more than two possible outcomes, and the second allows for varying levels of systematic bias at different levels of expertise. The aggregator is then evaluated under multiple synthetic data scenarios and on a large real-world dataset. The real data were collected by recruiting over 1300 forecasters, ranging from graduate students to forecasting and political science faculty and practitioners, and then posing them 69 geopolitical prediction problems (see the Appendix for a complete listing of the problems). Our main contribution arises from our ability to evaluate competing aggregators on the largest dataset ever collected on geopolitical probability forecasts made by human experts. Given such a large dataset, we have been able to develop a generic aggregator that is analytically simple and yet outperforms other widely used competing aggregators in practice. After presenting the evaluation results, the paper concludes by exploring some future research ideas.

2. Theory

Using the logit function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

a probability forecast $p \in [0, 1]$ can be mapped uniquely to a real number called the log-odds, $\text{logit}(p) \in \mathbb{R}$. This allows us to conveniently model probabilities with well-studied distributions, such as the Normal distribution, that are defined on the entire real line. In this section, assume that we have N experts who each provide *one* probability forecast of a binary-outcome event. We consider these experts to be interchangeable. That is, no one forecaster can be distinguished from the others either across or within problems. Denote the experts' forecasts by p_i and let $Y_i = \text{logit}(p_i)$ for $i = 1, 2, \ldots, N$. As was discussed earlier, we model the log-odds using a Normal distribution centered at the "true log-odds", which have been regressed towards zero by a factor of a. More specifically,

$$Y_i = \log\left(\frac{p}{1-p}\right)^{1/a} + \epsilon_i,$$

where $a \geq 1$ is an unknown level of systematic bias, p is the "true probability" to be estimated, and each $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ is a random shock with an unknown variance σ^2 on the individual's reported log-odds. If the model is correct, the event arising from this model would occur with

[1] In this paper, we use quotation marks in any reference to a true probability (or log-odds), in order to avoid a philosophical discussion. These quantities should be viewed simply as model parameters that are subject to estimation.

[2] The Brier score is the squared distance between the probability forecast and the event indicator that is equal to 1.0 or 0.0, depending on whether the event happened or not, respectively.

probability p. Therefore, p should be viewed as a model parameter that is subject to estimation.

The larger a is, the more the log-odds are regressed towards 0, or, equivalently, the more the probability estimates are regressed towards 0.5. We therefore associate $a = 1$ with an accurate forecast, and any $a > 1$ with a partially informed and under-confident forecast (Baron et al., submitted for publication). It is certainly possible for an expert to be overconfident (see for example McKenzie, Liersch, & Yaniv, 2008, for a recent and comprehensive discussion). In fact, we find this to be the case among forecasters at the highest level of self-reported expertise. On the other hand, we provide empirical evidence in Section 3.3.3 that, as a group, the forecasters tend to be under-confident. We therefore treat group-level under-confidence as a reasonable modeling restriction that we do not need to impose in our simulations (see Section 3), as we allow the data to speak for themselves by letting $a \in [0, \infty)$.

Note that, unlike the systematic bias term a, the random error component ϵ_i is allowed to vary among experts. Putting all of this together gives

$$\log\left(\frac{p_i}{1-p_i}\right) \overset{\text{i.i.d.}}{\sim} \text{Normal}\left(\log\left(\frac{p}{1-p}\right)^{1/a}, \sigma^2\right)$$

$$\Leftrightarrow \frac{p_i}{1-p_i} \overset{\text{i.i.d.}}{\sim} \text{Log-Normal}\left(\log\left(\frac{p}{1-p}\right)^{1/a}, \sigma^2\right)$$

$$\Leftrightarrow p_i \overset{\text{i.i.d.}}{\sim} \text{Logit-Normal}\left(\log\left(\frac{p}{1-p}\right)^{1/a}, \sigma^2\right).$$

This model is clearly based on an idealization of the real world, and is therefore an over-simplification. Although performing a formal statistical test to determine whether the log-odds in our real-world dataset follow a Normal distribution leads to a rejection of the null hypothesis of normality, this result simply reflects the inevitability of slight deviations from normality, and the sensitivity of the statistical tests when large sample sizes are involved. However, the assumption of normality turns out to be a good enough approximation to be of practical use. While Zhang and Maloney (2012) did not model log-odds using a Normal distribution, they argue in favor of using the logit-transformation with a linear bias term to model probabilities. Di Bacco, Frederic, and Lad (2003) use the Logit-Normal distribution to model experts' probabilities jointly under different levels of information. For our purposes, the Logit-Normal model serves as a theoretical basis for a clean and justified construction of an efficient aggregator.

2.1. Model-based aggregator

The invariance property of the maximum likelihood estimator (MLE) can be used to show that the MLE of p is

$$\hat{p}_G(a) = \frac{\exp\left(a\bar{Y}\right)}{1+\exp\left(a\bar{Y}\right)},$$

where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$. By plugging in the definition of Y_i, the MLE can be expressed in terms of the geometric mean of the odds as

$$\hat{p}_G(a) = \frac{\left[\prod_{i=1}^{N}\left(\frac{p_i}{1-p_i}\right)^{1/N}\right]^a}{1+\left[\prod_{i=1}^{N}\left(\frac{p_i}{1-p_i}\right)^{1/N}\right]^a}, \tag{1}$$

where the subindex G indicates the use of the geometric mean. The input argument emphasizes the dependency on the unknown quantity a. The estimator $\hat{p}_G$ is particularly convenient, because it allows for (i) an easy extension to uneven expert weights by simply replacing each $1/N$ with a weight term w_i, and (ii) switching the order of the transformation and aggregation operators. Note, however, that making use of (i) would result in an estimator with a total of N parameters. Such an estimator would be computationally complex and susceptible to overfitting. Many authors, including Armstrong (2001), Clemen (1989), and Graefe, Armstrong, Jones, and Cuzán (2014), have encouraged the use of equal weights unless there is strong evidence supporting unequal weightings of the experts. For the sake of simplicity, we limit this paper to the equally weighted aggregator.

2.2. Estimating systematic bias

Our aggregator $\hat{p}_G$ depends on the unknown quantity a, which needs to be inferred. If we have a training set consisting of K binary-outcome events and K pools of probability forecasts associated with these events, we can measure the goodness-of-fit for any a using the mean score

$$\bar{S}_K(a) = \frac{1}{K}\sum_{k=1}^{K} S(\hat{p}_{G,k}(a), Z_k),$$

where S is a proper scoring rule (see for example Gneiting & Raftery, 2007), $\hat{p}_{G,k}$ is the aggregate probability forecast for the kth event, and the event indicator $Z_k \in \{0, 1\}$ depends on whether the kth event occurred ($Z_k = 1$) or not ($Z_k = 0$). Optimizing this mean score as a function of a gives the *optimum score estimator*

$$\hat{a}_{OSE} = \arg\min_a \bar{S}_K(a),$$

which, according to Gneiting and Raftery (2007), is a consistent estimator of a.

Although strictly proper scoring rules are the natural loss functions when estimating binary class probabilities (see Buja, Stuetzle, & Shen, 2005), the real appeal arises from the freedom to choose a proper scoring rule to suit the problem at hand. Among the infinite number of proper scoring rules, the two most popular ones are the Brier score (see Brier, 1950) and the logarithmic scoring rule (see Good, 1952), which is equivalent to maximizing the log-likelihood, and hence finding the maximum likelihood estimator of a. Given that it is not clear which rule should be used for predicting social science events, we estimate a

via both the Brier score

$$\hat{a}_{BRI} = \arg\min_a \sum_{k=1}^{K} \left(\hat{p}_{G,k}(a) - Z_k\right)^2$$

and the likelihood function

$$\hat{a}_{MLE} = \arg\max_a \prod_{k=1}^{K} \hat{p}_{G,k}(a)^{Z_k}(1-\hat{p}_{G,k}(a))^{1-Z_k},$$

and compare the two resulting aggregators. Note that both equations are non-linear optimization problems with no analytical solutions. Fortunately, the optimizing values can be found using numerical methods such as the Newton–Raphson method, or by a simple line search.

2.3. Extensions to the aggregator

This section briefly discusses two extensions to the aggregator. The first extends $\hat{p}_G$ to events with more than two possible outcomes. This gives a more general aggregator, with $\hat{p}_G$ as a sub-case. The second allows for varying values of a across groups with different levels of expertise.

2.3.1. Multinomial events

For now, assume that the event can take exactly one of a total of $M \geq 2$ different outcomes. Under pure ignorance, the forecaster should assign a probability of $1/M$ to each outcome. The more ignorant the forecaster is, the more we would expect him to shrink his forecasts towards $1/M$. See Fox and Rottenstreich (2003) and Zhang and Maloney (2012) for further discussion.

We use this idea to generalize our aggregator. Choosing the Mth outcome as the baseline, denoting the forecast from the ith forecaster for the mth outcome by $p_{i,m}$, and letting $Y_{i,m} = \log\left(\frac{p_{i,m}}{p_{i,M}}\right)$ for $i = 1, 2, \ldots, N$, we arrive at a more general version of the model represented as

$$Y_{i,m} = \log\left(\frac{p_m}{p_M}\right)^{1/a} + \epsilon_{i,m},$$

where $m \in \{1, \ldots, M-1\}$ and $\epsilon_{i,m} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, with σ^2 unknown. The resulting maximum likelihood estimator for the kth outcome is

$$\hat{p}_{G,k}(a) = \frac{\left[\prod_{i=1}^{N}\left(\frac{p_{i,k}}{p_{i,M}}\right)^{1/N}\right]^a}{\sum_{j=1}^{M}\left[\prod_{i=1}^{N}\left(\frac{p_{i,j}}{p_{i,M}}\right)^{1/N}\right]^a}.$$

Instead of analyzing this more general estimator, this paper will focus on the binary sub-case. Note, however, that all of the properties generalize trivially to the multi-outcome case.

2.3.2. Correction under levels of expertise

The reasoning in the previous subsection suggests that a better forecast performance could be achieved by correcting for the systematic bias differently at different levels of expertise. To make this more specific, assume that each forecaster can identify himself with one of C levels of expertise, with C being the most knowledgeable. Let $\boldsymbol{a} = [a_1, \ldots, a_C]'$ be a vector of C different systematic bias factors, one for each expertise level. Then,

$$\hat{p}_{G,k}(\boldsymbol{a}) = \frac{\prod_{i=1}^{N}\left(\frac{p_{i,k}}{p_{i,M}}\right)^{\frac{\boldsymbol{e}_i'\boldsymbol{a}}{N}}}{\sum_{j=1}^{M}\prod_{i=1}^{N}\left(\frac{p_{i,j}}{p_{i,M}}\right)^{\frac{\boldsymbol{e}_i'\boldsymbol{a}}{N}}},$$

where $\boldsymbol{e}_i$ is a vector of length C, indicating which level of expertise the ith forecaster belongs to. For instance, if $\boldsymbol{e}_i = [0, 1, 0, \ldots, 0, 0]'$, the ith expert identifies himself with expertise level two. The systematic bias factors can be estimated by first partitioning the dataset by expertise, then finding the optimal value for each expert group separately. We will return to this topic briefly at the end of the results section, where we show the effects of forecasters' actual ratings of their own expertise.

3. Results and discussion

This section compares different aggregators on both synthetic and real-world data. The aggregators included in the analysis are as follows.

(a) Arithmetic mean of the probabilities.

(b) Median of the probabilities.

(c) Logarithmic opinion pool

$$\hat{p} = \prod_{i=1}^{N} p_i^{w_i} \Bigg/ \left(\prod_{i=1}^{N} p_i^{w_i} + \prod_{i=1}^{N}(1-p_i)^{w_i}\right),$$

which, according to Bacharach (1972), was proposed by Peter Hammond (see Genest & Zidek, 1986). Given that we consider the forecasters to be indistinguishable, we assign equal weights to each forecaster. Letting $w_i = 1/N$ for $i = 1, \ldots, N$ gives us the equally weighted logarithmic opinion pool (ELOP).

(d) Our aggregator $\hat{p}_G(a)$ as given by Eq. (1).

(e) The Beta-transformed linear opinion pool

$$\hat{p}(\alpha, \beta) = H_{\alpha,\beta}\left(\sum_{i=1}^{N} w_i p_i\right),$$

where $H_{\alpha,\beta}$ is the cumulative distribution function of the Beta distribution with parameters α and β, and w_i is the weight given to the ith forecast. Using simulations, Allard et al. (2012) show that Beta-transformed linear pooling presents a very good forecast performance. Again, we assign equal weights to each forecaster. Letting $w_i = 1/N$ for $i = 1, \ldots, N$ gives us the Beta-transformed equally weighted linear opinion pool (BELP). However, this aggregator tends to overfit in all of our evaluation procedures, and a much better performance can be obtained by requiring $\alpha = \beta \geq 1$. Under such a restriction, the BELP aggregator can be required to shift any mean probability more toward the closest extreme probability, 0.0 or 1.0. This one-parameter sub-case (1P-BELP) is more robust against overfitting, and is supported by the theoretical results of Wallsten and Diederich (2001). For these reasons, it is a good competing aggregator in our

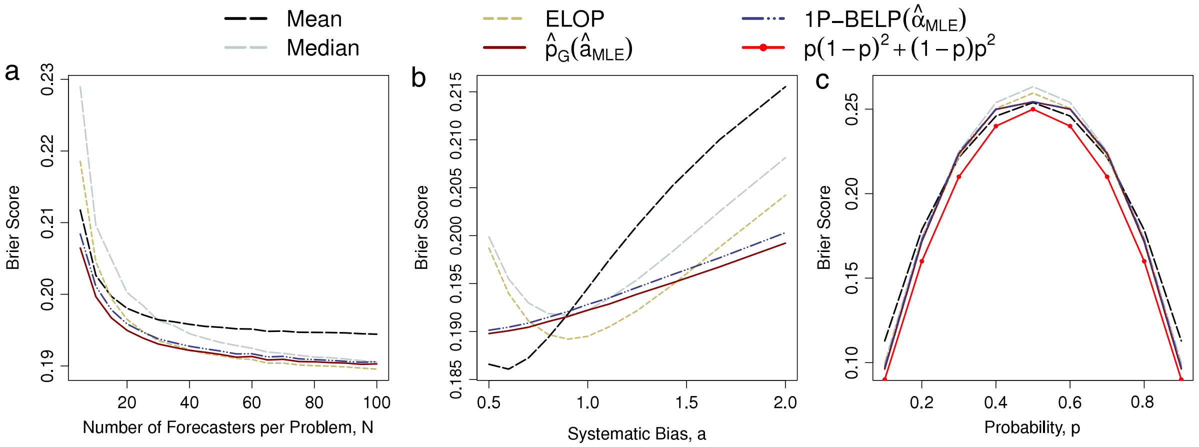

Fig. 1. $K = 30$ synthetic problems for training; 1000 problems for testing.

simulations. We do not present the results associated with the 2-parameter BELP aggregator because BELP performs much worse than 1P-BELP in all of our simulations.

As was suggested by Ranjan and Gneiting (2010), the parameter α can be fitted using optimum score techniques. We fitted any tuning parameters using both the Brier score and the likelihood function, and then compared the resulting aggregators. Given that Ranjan and Gneiting (2010) only considered aggregating binary events, it is not clear how the Beta-transformed linear pooling can be generalized to events with more than two possible outcomes. Therefore, our comparison will focus only on forecasting binary events.

Throughout this evaluation section, we will be using the Brier score as the measure of performance. As was discussed earlier in Section 2.2, this scoring rule has some attractive properties, and is, in essence, a quadratic penalty. It also has an interesting psychological interpretation as the expected cost of an error, given a probability judgment and the truth (see Baron et al., submitted for publication, for details).

3.1. Synthetic data: correctly specified model

In this section, we evaluate the different aggregators on a correctly specified model; that is, on data that have been generated directly from the Logit-Normal distribution described in Section 2. The evaluation is done over a three-dimensional grid that expands the number of forecasters per problem, N, from 5 to 100 (in increments of five forecasters), the true probability, p, from 0.1 to 0.9 (in increments of 0.1), and the systematic bias term, a, from $5/10, 6/10, \ldots, 9/10, 10/10, 10/9, \ldots, 10/6, 10/5$ symmetrically around the no-bias point at 1.0. The simulation was run for 100 iterations at every grid point. Each iteration used the values at the grid point to produce a synthetic data set from the Logit-Normal distribution. The true probability, p, was used to generate Bernoulli random variables that indicated which events occurred and which did not. Testing was performed on a separate testing set consisting of 1000 problems, each with the same number of forecasters as the problems in the original training set. The simulation was repeated for two different numbers of problems in the training set, $K = 30$ and $K = 100$. The variance for the log-odds, σ^2, was equal to five throughout the entire simulation.[3]

The results are summarized in two sets of figures: Figs. 1(a) and 2(a) plot the Brier scores (given by averaging over the systematic bias and the true probability) against the number of forecasters per problem; Figs. 1(b) and 2(b) plot the Brier scores (given by averaging over the number of forecasters per problem and the true probability) against the systematic bias term; and Figs. 1(c) and 2(c) plot the Brier scores (given by averaging over the number of forecasters per problem and the systematic bias) against the true probability. Fig. 1 presents the results under $K = 30$, and Fig. 2 shows the results under $K = 100$.

Given that $\hat{p}_G(\hat{a}_{MLE})$ and 1P-BELP $(\hat{\alpha}_{MLE})$ performed better than $\hat{p}_G(\hat{a}_{BRI})$ and 1P-BELP $(\hat{\alpha}_{BRI})$, only the results associated with the maximum likelihood approach are presented. Comparing Figs. 1 and 2 shows that these two aggregators make very good use of the training data, and outperform the simple, parameterless aggregators as the training set increases from $K = 30$ to $K = 100$ problems. Overall, our aggregator $\hat{p}_G(\hat{a}_{MLE})$ achieves the lowest Brier score almost uniformly across Fig. 2(a)–(c). However, this result is more of a sanity-check than a surprising result, as the data were generated explicitly to match the model assumptions made by $\hat{p}_G$.

Based on Figs. 1(b) and 2(b), correcting for the bias when the data are actually unbiased ($a = 1.0$) does not cause much harm, but correcting for the bias when the data are truly biased ($a \neq 1.0$) yields noticeable performance benefits, especially when $K = 100$. Interestingly, the mean performs better than any of the other aggregators when the experts are highly over-confident ($a \leq 0.7$), but is hugely

[3] This value was considered a realistic choice after analyzing the variance of the log-odds in our real-world data. The simulation was also run with a unit variance. However, these results were not remarkably different, and hence, are not presented in this paper, for the sake of brevity.

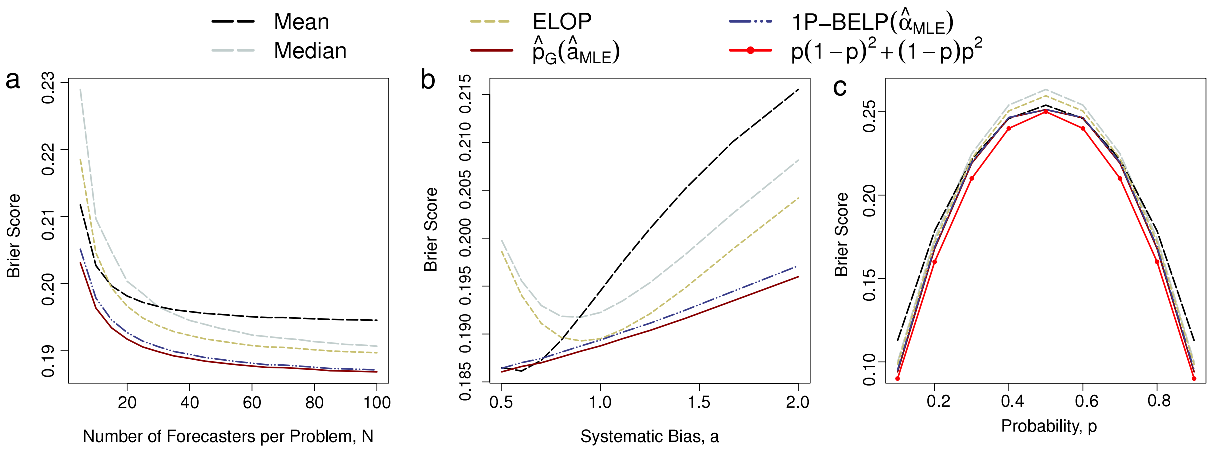

Fig. 2. $K = 100$ synthetic problems for training; 1000 problems for testing.

outperformed when the experts are under-confident ($a > 1$). In order to gain some understanding of this behavior, notice that, in the highly over-confident case, the distribution of the forecasts tends to be very skewed in the probability scale. The values in the long tail of such a distribution have a larger influence on the mean than, say, the median of the probability forecasts. Given that the median remains mostly unaffected by these values, it produces an aggregate forecast that remains over-confident. In contrast, the mean is drawn towards 0.5 by the values in the long tail. This produces an aggregate forecast that is less over-confident, hence improving the forecast performance.

However, this improved performance comes at a cost: when the true probability p is very close to the extreme probabilities 0.0 and 1.0, the mean is, on average, the worst performer among all of the aggregators in the analysis. This difference in performance, which is clear in Figs. 1(c) and 2(c), is more meaningful when it is compared to the baseline given by $p(1-p)^2 + (1-p)p^2$. Given that the expected Brier score is minimized at the true probability, this line should be considered as the ultimate goal in Figs. 1(c) and 2(c). Note that all aggregators, except the mean, approach the line $p(1-p)^2 + (1-p)p^2$ from above as p gets closer to the extreme probabilities 0.0 and 1.0.

3.2. Synthetic data: misspecified model

Next, we evaluate the different aggregators on data that have not been generated from the Logit-Normal distribution. The setup considered is an extension of the simulation study introduced by Ranjan and Gneiting (2010) and further applied by Allard et al. (2012). Under our extended version, the true probability for a problem with N forecasters is given by

$$p = \Phi\left(\sum_{i=1}^{N} u_i\right),$$

where Φ is the cumulative distribution function of a standard normal distribution, and each $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Now, assume that the ith expert is aware of the true probability generating process but only observes u_i. Then, his calibrated estimate for p is given by

$$p_i = \Phi\left(\frac{u_i}{\sqrt{2N-1}}\right).$$

Note that the larger the number of forecasters participating in a given problem, the less information (proportionally) knowing u_i gives the forecaster. Therefore, as the number of forecasters increases, the forecaster will shrink his estimate more and more towards 0.5. More specifically, $p_i \to \Phi(0) = 0.5$ for all $i = 1, \ldots, N$ as $N \to \infty$.

For a real-world analogy of this setup, think of a group of N people who are voting independently on a binary event. Knowing everybody's vote would determine the final outcome. Given that each person only knows his own vote, his proportional knowledge share diminishes as more people enter the voting. As a result, his probability forecast for the final outcome should shrink towards 0.5.

In our simulation, we varied the number of forecasters per problem, N, between 2 and 100 (in increments of one forecaster). Under each value of N, the simulation ran for a total of 10,000 iterations. Each iteration produced the true probabilities for the K problems and their associated pools of N probability estimates from the process described above. The true probabilities were then used to generate Bernoulli random variables that indicated which events occurred and which did not. Testing was performed on a separate testing set consisting of 1000 problems, each with the same number of forecasters as the problems in the training set. In the end, the resulting Brier scores were averaged to give an average Brier score for each number of forecasters for each problem.

Fig. 3 plots the average Brier score against the number of forecasters per problem under $K = 100$ problems. The same analysis was also performed under $K = 30$, but the results turned out to be almost identical to those under $K = 100$, and hence, they are not presented separately here, for the sake of brevity. Before discussing the $K = 100$ results, however, it is important to emphasize the peculiarity of this setting. Note that, unlike in many commonly encountered data generating processes, having more data

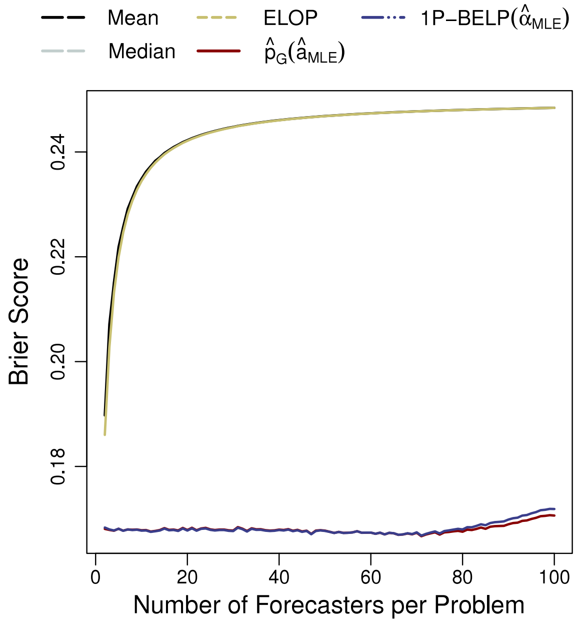

Fig. 3. 100 synthetic problems for training; 1000 problems for testing.

leads to an increased bias, and is therefore harmful. As a result, we would expect the aggregators to perform worse as the sample size increases. Based on Fig. 3, the mean, median, and ELOP, which do not aim to correct for the bias, do in fact degrade in terms of their performances as the number of forecasters increases. In contrast, the one-parameter aggregators, $\hat{p}_G$ and 1P-BELP, are able to stabilize the average Brier score, despite the increasing bias in the probability estimates. Overall, $\hat{p}_G$ achieves the lowest Brier scores across all numbers of forecasters per problem.

3.3. Real data: predicting geopolitical events

We recruited over 1300 forecasters, who ranged from graduate students to forecasting and political science faculty and practitioners, and then asked them to provide probability forecasts for 69 geopolitical events. The forecasters were recruited from professional societies, research centers, alumni associations and science bloggers, as well as by word of mouth. The requirements included at least a Bachelor's degree and the completion of psychological and political tests that took roughly two hours. These measures assessed cognitive styles, cognitive abilities, personality traits, political attitudes, and real-world knowledge. All of the forecasters knew that their probability estimates would be assessed for accuracy using Brier scores. This gave the forecasters incentives to report their true beliefs instead of trying to game the system. Each forecaster received $150 for meeting certain minimum participation requirements, regardless of their accuracy. They also received status rewards for their performances via leaderboards displaying Brier scores for the top 20 forecasters. Each of the 69 geopolitical events had two possible outcomes. For instance, two of the questions were

> *Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?*

and

> *Will the Nikkei 225 index finish trading at or above 9500 on 30 September 2011?*

See the Appendix for a complete list of the 69 problems and the associated summary statistics.

The forecasters were allowed to update their forecasts as long as a question was active. Some questions were active longer than others. The number of days active ranged from 2 to 173 days, with a mean of 54.7 days. It is important to note that this paper does not focus on dynamic data. Instead, we study pools of probability forecasts, with no more than one forecast given by a single expert. More specifically, we consider the first three days for each problem, because this is when the highest level of uncertainty is present. If an expert made more than one forecast during these three days, we consider only his most recent forecast. This results in 69 sets of probabilities, with the number of forecasters per problem ranging from 86 to 647, with a mean of 235.1. Given that not all experts participated in every problem, we consider the experts as being completely anonymous (and interchangeable) both within and across problems. However, before we evaluate the results, we discuss several practical matters that need to be taken into account when aggregating real-world forecasting data.

3.3.1. Extreme values and inconsistent data

For any value of a, the aggregator $\hat{p}_G$ satisfies the 0/1 forcing property, which states that if the pool of forecasts includes an extreme value, that is, either zero or one but not both, then the estimator should return that extreme value (see for example Allard et al., 2012). This property is desirable if one of the forecasters happens to know the final outcome of the event with absolute confidence. Unfortunately, experts sometimes make such absolute claims even when they are not completely sure of the outcome. For instance, the forecast pools associated with each of the 69 questions in our data all contained both a zero and a one. In any such dataset, an aggregator that is based on the geometric mean of the odds is undefined.

These data inconsistencies can be avoided by adding (subtracting) a small quantity to (from) zeros (ones). Ariely et al. (2000) suggest changing $p = 0$ to $p = 0.02$ and $p = 1$ to $p = 0.98$. Allard et al. (2012) only consider probabilities that fall within a constrained interval, say [0.001, 0.999], and throw out the rest. Given that this implies ignoring a portion of the data, we take an approach similar to that of Ariely et al. (2000), and replace $p = 0$ and 1 with $p = 0.01$ and 0.99, respectively. In the case of multinomial events, the modified probabilities should be normalized to sum to one. This forces the probability estimates to the open interval (0, 1). The transformation will shift the truncated values even closer to their true extreme values. For instance, if a is larger than two, as is often the case, 0.01 and 0.99 would be transformed to at least 0.0001 and 0.9999, respectively.

Another practical solution is to estimate the geometric mean of the odds based on the odds given by the arithmetic mean of the probabilities. This gives us the following estimator:

$$\hat{p}_A(a) = \frac{\left[\frac{\bar{p}}{1-\bar{p}}\right]^a}{1+\left[\frac{\bar{p}}{1-\bar{p}}\right]^a},$$

where $\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$. The subindex emphasizes the use of the arithmetic mean. The two estimators $\hat{p}_G$ and $\hat{p}_A$ will differ the most when the set of probability forecasts includes values which are close to the extremes. Therefore, the larger the variance term σ^2 of the Logit-Normal model is, the more we would expect these two estimators to differ. For comparison's sake, we have included $\hat{p}_A$ in the real-world data analysis.

A similar problem arises with the logarithmic opinion pool, where zero predictions from experts can be viewed as "vetoes" (see Genest & Zidek, 1986). To address this, we replaced $p = 0$ with $p = 0.01$ and normalized the probabilities to sum to one.

3.3.2. Aggregator comparison on expert data

This section evaluates the aggregators on the first three days of the 69 problems in our dataset. The evaluation begins by exploring the impact of the number of forecasters per problem on the predictive power. Each run of the simulation fixes the number of forecasters per problem, and samples a random subset (of this size) of forecasters within each problem. These subsets are then used for training and computing a Brier score. The sampling procedure is repeated 1000 times during the simulation. The resulting 1000 Brier scores are then averaged to obtain an overall performance measure for the given number of forecasters per problem.

Fig. 4 plots the average Brier score against the number of forecasters per problem. The MLE aggregator $\hat{p}_G$ achieves the lowest Brier score across all numbers of forecasters per problem. The two aggregators $\hat{p}_A$ and P1-BELP perform so similarly that their average Brier scores are almost indistinguishable. The performance gap between $\hat{p}_G$ and P1-BELP and $\hat{p}_A$ appears to widen as the number of forecasters increases.

It is worth noting that most of the improvements across the different approaches occur before roughly 50 forecasters per problem. This suggests a new strategy for collecting data: instead of having a large number of forecasters making predictions on a few problems, we should have around 50 forecasters involved with a large number of problems. With a larger number of problems, more accurate estimates of the systematic bias could be acquired, possibly leading to improved forecast performances.

A similar analysis was performed using the last three days of each problem, but the average Brier scores were very close to zero. In fact, there was so much certainty among the forecasters, that simply taking the median gave an aggregate forecast which was very close to the truth. For this reason, we decided to not present these results in this paper.

The average Brier scores in Fig. 4 are based on the training error. No separate testing set was used in this particular analysis, because we believe that fitting one parameter in a large enough sample will not overfit significantly. Fig. 5 plots the Brier score for $\hat{p}_G$ under varying levels of a. Given that the optimal level of a is around 2.0, the experts (as a group) appear under-confident, and $\hat{p}_G$ gains its advantage by shifting each of the probability forecasts closer to its nearest boundary point (0.0 or 1.0).

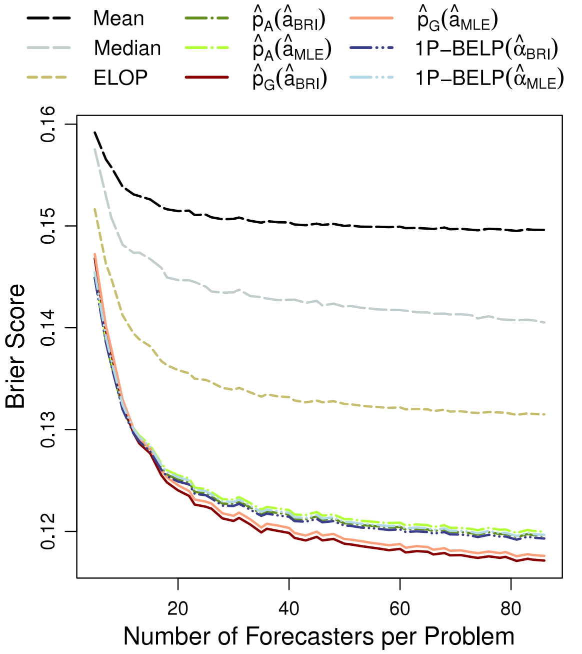

Fig. 4. 69 real-world problems for training. The first three days; much uncertainty.

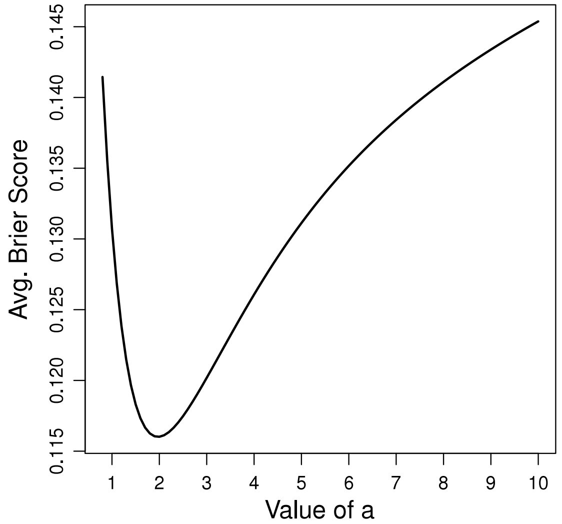

Fig. 5. Sensitivity to the choice of *a* based on all data available in the first three days.

Running a *repeated sub-sampling validation* with a training set of size K and a testing set of size $69 - K$ supports the results shown in Fig. 4. Table 1 shows the results after running *repeated sub-sampling validation* with $K = 30$ and $K = 60$ a total of 1000 times, and then averaging the resulting 1000 (testing) Brier scores. For the sake of

Table 1
K problems for training; $69-K$ problems for testing; 1000 repetitions. The values in parentheses are the estimated standard deviations of the testing scores.

	Brier score		Logarithmic score	
	Bias correction		Bias correction	
	No	Yes	No	Yes
Mean	0.150 (0.032)	0.150 (0.032)	0.477 (0.025)	0.477 (0.025)
Median	0.140 (0.038)	0.139 (0.038)	0.446 (0.031)	0.444 (0.031)
ELOP	**0.132** (**0.039**)	**0.131** (**0.039**)	**0.425** (**0.032**)	**0.425** (**0.032**)
	$K=30$			
	Bias estimation		Bias estimation	
	BRI	MLE	BRI	MLE
1P-BELP	0.126 (0.027)	0.125 (0.026)	0.401 (0.115)	0.401 (0.117)
$\hat{p}_A$	0.127 (0.027)	0.125 (0.026)	0.402 (0.109)	0.402 (0.115)
$\hat{p}_G$	**0.124** (**0.028**)	**0.122** (**0.026**)	**0.401** (**0.134**)	**0.394** (**0.127**)
	$K=60$			
	Bias estimation		Bias estimation	
	BRI	MLE	BRI	MLE
1P-BELP	0.122 (0.061)	0.121 (0.064)	0.383 (0.170)	0.384 (0.193)
$\hat{p}_A$	0.122 (0.061)	0.121 (0.065)	0.385 (0.168)	0.386 (0.190)
$\hat{p}_G$	**0.119** (**0.060**)	**0.118** (**0.064**)	**0.376** (**0.165**)	**0.377** (**0.188**)

consistency, we have also included the average logarithmic scores:

$$-\frac{1}{69-K}\sum_{k=1}^{69-K} Z_k \log(\hat{p}_k) + (1-Z_k)\log(1-\hat{p}_k),$$

where $\hat{p}_j$ is the probability estimate and Z_j is the event indicator for the jth testing problem, defined earlier in Section 2.2. The values given in parentheses are the estimated standard deviations of the testing scores. Given that the mean, median, and ELOP do not use training data, their reported scores are based on the simulation with $K = 30$ that uses a larger testing set.

In Table 1, we have also included the bias-corrected versions of the mean, median, and ELOP. This correction was obtained by applying bootstrap sampling to the pool of probabilities a total of 1000 times. More specifically,

$$\hat{p}_{f,k} = 2f(\boldsymbol{p}_k) - \frac{1}{1000}\sum_{i=1}^{1000} f\left(\boldsymbol{p}_{k,bs}^{(i)}\right),$$

where $\boldsymbol{p}_k$ is the (full) original set of probabilities for the kth problem, $\boldsymbol{p}_{k,bs}^{(i)}$ is the ith bootstrap sample obtained from $\boldsymbol{p}_k$, and f is a functional which depends on the estimator. For instance, when correcting the sample median, $f(\boldsymbol{p}_k) = \text{median}(\boldsymbol{p}_k)$. However, the biases found turned out to be very small. As can be seen from Table 1, correcting for the bias only improved the performance by a small margin, if at all.

For convenience, we have put the lowest scores in each column of the three boxes in bold. Overall, the ranking of the aggregators based on relative performances is the same as in Fig. 4. As was seen earlier, $\hat{p}_G(\hat{a}_{MLE})$ achieves the lowest Brier and logarithmic scores by a noticeable margin.

3.3.3. Less transformation for more expertise

Earlier, we proposed that the more expertise the forecaster has, the less systematic bias will be found in his probability forecasts. This means that his forecasts require less of a transformation, i.e. a lower level of a. To evaluate this interpretation, we asked forecasters to self-assess their level of expertise on the topic. The level of expertise was measured on a scale of 1–5 ($1 =$ not at all expert to $5 =$ extremely expert). Fig. 6 plots the maximum likelihood estimator of a under different levels of expertise. The gray bars around each point are the 95% (Bonferroni corrected) simultaneous confidence intervals computed by inverting the likelihood-ratio test. We have also allowed for values of a of less than 1, in order to reveal possible overconfidence.

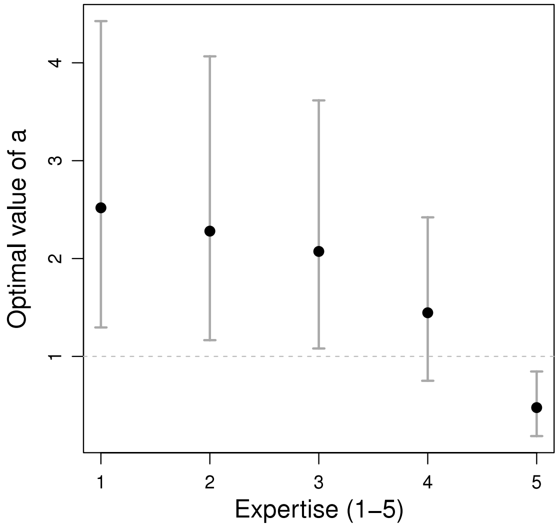

Fig. 6. Optimal transformation a (representing the systematic bias), with 95% simultaneous confidence intervals as a function of forecaster self-reported expertise.

These results are based on the first three days of data for each problem, because this is when the most uncertainty is present and the expertise level is most important. Although we are unable to show statistical significance for a strictly decreasing trend in the systematic bias across the different levels of expertise, the need for transformation is apparent in the 99% confidence intervals for the value of a when the level of expertise is not taken into account. This interval (not shown on Fig. 6) is [1.161, 3.921]. Given that it does not include $a = 1$, i.e., the level of no transformation, there is significant evidence (at the 1% significance level) that, as a group, the experts are under-confident. Therefore, their probability forecasts should be shifted more toward the extreme probabilities 0.0 and 1.0.

4. Conclusions

In this paper we have motivated and derived a model-based approach to the aggregation of expert probability forecasts. The resulting aggregator, which is based on the geometric mean of the expert odds, has a single tuning parameter that determines how much each of the probabilities should be shifted toward its nearest extreme probability, 0.0 or 1.0. This transformation aims to compensate for the under-confidence that arises from incomplete knowledge and tends to be present among experts at a group level. That is, although the individual experts may not all be under-confident (in fact, according to our analysis, some of the experts with high self-reported levels of expertise tend to be over-confident), as a group the experts are under-confident. Therefore, if no bias-correction is performed, the consensus probability forecast can turn out to be biased and sub-optimal in terms of forecast performance.

In studying the extent of this bias, it is helpful to compare the aggregate probability forecasts given by a naïve approach, such as the arithmetic mean with no explicit bias-correction, with the corresponding forecasts given by a bias-correcting approach, such as our $\hat{p}_G(\hat{a}_{MLE})$ aggregator. In the table in the Appendix, we have provided both the mean probability forecast and the aggregate estimate $\hat{p}_G(\hat{a}_{MLE})$ for the 69 problems in our real-world dataset. Looking at these estimates (see Fig. 7 in the Appendix), it is clear that the $\hat{p}_G(\hat{a}_{MLE})$ aggregator is much sharper than the simple arithmetic mean. Furthermore, the noticeable disagreement between the two estimates (with a mean absolute difference of 0.175) suggests that a large enough bias persists for bias-correction to improve the performance.

As is evident throughout Section 3, our aggregator shows a very good forecast performance, especially when the outcome of the event involves much uncertainty. In addition, our aggregator utilizes the training data efficiently, leading to an improved forecast performance as the size of the training set increases. However, this improvement happens at such a diminishing rate that there are few additional gains in forecast performance from aggregating more than about 50 forecasters per problem (see Fig. 4).

It is likely that our aggregator could be improved and extended in many ways. However, this might lead to a reduced interpretability and to additional assumptions that may not comply with the psychology literature. For instance, being able to estimate the bias term a within each problem individually could improve the performance of the aggregator. However, this seems difficult, given the framework of this paper; that is, non-dynamic probability pools given by interchangeable forecasters. As Table 1 shows, simple bootstrap approaches to problem-specific bias-corrections do not lead to significant improvements in forecast performance.

Perhaps an intermediate approach that neither shares a single bias term nor has completely independent bias terms across problems will yield further improvements in performance. One possibility is that the more difficult the problem, the more the bias will persist among the experts. This suggests that better predictions could be achieved by developing a measure of the difficulty of the problem, estimating a single bias term across all problems, and then adjusting this bias term individually for each problem, based on the estimated difficulty. However, coming up with a reasonable difficulty measure is challenging. One simple idea would be to use the variance of the expert forecasts as a proxy for problem difficulty.

Such an extension could also satisfy the unanimity property: if all experts give the same forecast, then the aggregator should return that forecast as a unanimous decision. Although this property may not be critical in large probability pools like our dataset, it needs to be mentioned that our aggregator does not satisfy the unanimity property. Instead, it tends to assume under-confidence and shift each of the probability forecasts closer to its nearest extreme probability, 0.0 or 1.0. Nonetheless, it gives extremely good results on real data. Furthermore, unlike the Beta-transformed linear opinion pool, our aggregator can be applied to a wide range of situations, such as events with more than two possible outcomes, and has a simple analytical form, making it interpretable, flexible, and amenable to many future extensions.

Acknowledgments

This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

We deeply appreciate the project management skills and work of Terry Murray and David Wayrynen, which went far beyond the call-of-duty on this project.

Appendix

Unfortunately, the full real-world dataset is not accessible to the public at the moment. However, we have requested permission to publish the data online in the near future. For the time being, we have included the following table that shows a complete list of the 69 problems in our dataset. For each problem, six summary statistics have

Table A.1
Summary of the 69 problems in our dataset.

Question text	$\hat{p}_G$	$\bar{p}$	s_p	N	T	Z
Will the Six-Party talks (among the US, North Korea, South Korea, Russia, China, and Japan) formally resume in 2011?	0.04	0.27	0.22	102	123	0
Will Serbia be officially granted EU candidacy by 31 December 2011?	0.03	0.27	0.24	96	124	0
Will the United Nations General Assembly recognize a Palestinian state by 30 September 2011?	0.01	0.23	0.27	128	29	0
Will Daniel Ortega win another term as President of Nicaragua during the late 2011 elections?	0.78	0.61	0.19	109	65	1
Will Italy restructure or default on its debt by 31 December 2011?	0.28	0.44	0.26	86	124	0
By 31 December 2011, will the World Trade Organization General Council or Ministerial Conference approve the 'accession package' for WTO membership for Russia?	0.43	0.48	0.21	98	106	1
Will the 30 Sept 2011 "last" PPB for Nov 2011 Brent Crude oil futures exceed $115?	0.23	0.40	0.21	302	23	0
Will the Nikkei 225 index finish trading at or above 9500 on 30 September 2011?	0.06	0.29	0.21	290	22	0
Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, OR otherwise vacate office before 1 October 2011?	0.03	0.24	0.20	333	23	0
Will the London Gold Market Fixing price of gold (USD per ounce) exceed $1850 on 30 September 2011 (10 am ET)?	0.78	0.60	0.22	269	23	0
Will Israel's ambassador be formally invited to return to Turkey by 30 September 2011?	0.02	0.22	0.19	334	23	0
Will PM Donald Tusk's Civic Platform Party win more seats than any other party in the October 2011 Polish parliamentary elections?	0.80	0.61	0.19	281	31	1
Will Robert Mugabe cease to be President of Zimbabwe by 30 September 2011?	0.01	0.16	0.20	358	23	0
Will Muqtada al-Sadr formally withdraw support for the current Iraqi government of Nouri al-Maliki by 30 September 2011?	0.08	0.30	0.19	282	23	0
Will peace talks between Israel and Palestine formally resume at some point between 3 October 2011 and 1 November 2011?	0.02	0.23	0.21	309	28	0
Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?	0.64	0.55	0.26	395	9	1
Will the South African government grant the Dalai Lama a visa before 7 October 2011?	0.03	0.28	0.25	647	2	0
Will former Ukrainian Prime Minister Yulia Tymoshenko be found guilty on any charges in a Ukrainian court before 1 November 2011?	0.50	0.51	0.20	364	6	1
Will Abdoulaye Wade win re-election as President of Senegal?	0.79	0.62	0.16	200	173	0
Will the Freedom and Justice Party win at least 20% of the seats in the first People's Assembly (Majlis al-Sha'b) election in post-Mubarak Egypt?	0.86	0.65	0.19	207	108	1
Will Joseph Kabila remain president of the Democratic Republic of the Congo through 31 January 2012?	0.93	0.72	0.16	166	119	1
Will Moody's issue a new downgrade of the sovereign debt rating of the Government of Greece between 3 October 2011 and 30 November 2011?	0.83	0.64	0.22	203	57	0
Will the UN Security Council pass a measure/resolution concerning Syria in October 2011?	0.11	0.35	0.24	231	27	0
Will the US Congress pass a joint resolution of disapproval in October 2011 concerning the proposed $5+ billion F-16 fleet upgrade deal with Taiwan?	0.02	0.23	0.22	297	17	0
Will the Japanese government formally announce the decision to buy at least 40 new jet fighters by 30 November 2011?	0.30	0.44	0.20	193	57	0
Will the Tunisian Ennahda party officially announce the formation of an interim coalition government by 15 November 2011?	0.70	0.57	0.23	508	7	0
Will Japan officially become a member of the Trans-Pacific Partnership before 1 March 2012?	0.47	0.49	0.22	150	113	0
Will the United Nations Security Council pass a new resolution concerning Iran by 1 April 2012?	0.59	0.53	0.27	193	145	0
Will Hamad bin Isa al-Khalifa remain King of Bahrain through 31 January 2012?	0.99	0.82	0.17	163	84	1
Will Bashar al-Assad remain President of Syria through 31 January 2012?	0.95	0.71	0.24	143	84	1
Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, OR otherwise vacate office before 1 January 2012?	1.00	0.83	0.22	523	4	1
Will Lucas Papademos be the next Prime Minister of Greece?	0.94	0.70	0.24	388	2	1
Will Lucas Papademos resign, lose re-election/confidence vote, or vacate the office of Prime Minister of Greece before 1 March 2012?	0.17	0.38	0.24	231	79	0
Will the United Kingdom's Tehran embassy officially reopen by 29 February 2012?	0.02	0.21	0.20	237	79	0
Will a trial for Saif al-Islam Gaddafi begin in any venue by 31 March 2012?	0.41	0.48	0.26	215	110	0
Will S&P downgrade the AAA long-term credit rating of the European Financial Stability Facility (EFSF) by 30 March 2012?	0.69	0.57	0.22	259	33	1
Will North Korea successfully detonate a nuclear weapon, either atmospherically, underground, or underwater, between 9 January 2012 and 1 April 2012?	0.02	0.22	0.22	215	83	0
By 1 April 2012, will Egypt officially announce its withdrawal from its 1979 peace treaty with Israel?	0.01	0.18	0.19	227	83	0
Will Kim Jong-un attend an official, in-person meeting with any G8 head of government before 1 April 2012?	0.02	0.21	0.22	238	82	0
Will Christian Wulff resign or vacate the office of President of Germany before 1 April 2012?	0.16	0.37	0.23	241	38	1
Will the daily Europe Brent Crude FOB spot price per barrel be greater than or equal to $150 before 3 April 2012?	0.04	0.27	0.23	206	84	0
Will the Taliban begin official in-person negotiations with either the US or Afghan government by 1 April 2012?	0.08	0.32	0.24	184	69	0
Will Yousaf Raza Gillani resign, lose confidence vote, or vacate the office of Prime Minister of Pakistan before 1 April 2012?	0.18	0.38	0.22	146	68	0

Table A.1 (*continued*)

Question text	$\hat{p}_G$	$\bar{p}$	s_p	N	T	Z
Will Yemen's next presidential election commence before 1 April 2012?	0.34	0.46	0.25	222	28	1
Will Traian Basescu resign, lose referendum vote, or vacate the office of President of Romania before 1 April 2012?	0.08	0.31	0.21	149	68	0
Will the UN Security Council pass a new measure/resolution directly concerning Syria between 23 January 2012 and 31 March 2012?	0.39	0.47	0.26	156	68	0
Before 1 April 2012, will South Korea officially announce a policy of reducing Iranian oil imports in 2012?	0.46	0.49	0.24	170	68	0
Will Israel release Palestinian politician Aziz Duwaik from prison before 1 March 2012?	0.05	0.29	0.23	210	37	0
Will Iran and the US commence official nuclear program talks before 1 April 2012?	0.01	0.17	0.20	225	61	0
Will Serbia be officially granted EU candidacy before 1 April 2012?	0.07	0.30	0.23	253	31	1
Will the IMF officially announce before 1 April 2012 that an agreement has been reached to lend Hungary an additional 15+ Billion Euros?	0.51	0.51	0.23	177	61	0
Will Libyan government forces regain control of the city of Bani Walid before 6 February 2012?	0.18	0.38	0.24	500	6	0
Will a run-off be required in the 2012 Russian presidential election?	0.04	0.29	0.25	277	34	0
Will the Iraqi government officially announce before 1 April 2012 that it has dropped all criminal charges against its VP Tareq al-Hashemi?	0.04	0.27	0.22	200	61	0
Will Egypt officially announce by 15 February 2012 that it is lifting its travel ban on Americans currently in Egypt?	0.44	0.50	0.25	321	16	0
Will a Japanese whaling ship enter Australia's territorial waters between 7 February 2012 and 10 April 2012?	0.17	0.37	0.27	213	63	0
Will William Ruto cease to be a candidate for President of Kenya before 10 April 2012?	0.16	0.37	0.25	192	62	0
Will Marine LePen cease to be a candidate for President of France before 10 April 2012?	0.04	0.26	0.22	214	62	0
Between 21 February 2012 and 1 April 2012, will the UN Security Council announce any reduction of its peacekeeping force in Haiti?	0.20	0.41	0.25	168	40	0
Will Mohamed Waheed Hussain Manik resign or otherwise vacate the office of President of Maldives before 10 April 2012?	0.08	0.32	0.23	155	48	0
Will Japan commence parliamentary elections before 1 April 2012?	0.04	0.28	0.23	182	39	0
Before 13 April 2012, will the Turkish government officially announce that the Turkish ambassador to France has been recalled?	0.06	0.29	0.22	143	51	0
Will Standard and Poor's downgrade Japan's Foreign Long Term credit rating at any point between 21 February 2012 and 1 April 2012?	0.08	0.32	0.25	172	40	0
Will Myanmar release at least 100 more political prisoners between 21 February 2012 and 1 April 2012?	0.55	0.52	0.25	170	40	0
Will a civil war break out in Syria between 21 February 2012 and 1 April 2012?	0.54	0.51	0.24	191	40	0
Will Tunisia officially announce an extension of its current state of emergency before 1 April 2012?	0.77	0.61	0.24	198	26	1
Before 1 April 2012, will Al-Saadi Gaddafi be extradited to Libya?	0.02	0.22	0.20	225	26	0
Before 1 April 2012, will the Sudan and South Sudan governments officially announce an agreement on oil transit fees?	0.04	0.27	0.23	202	26	0
Will Yemeni government forces regain control of the towns of Jaar and Zinjibar from Al-Qaida in the Arabian Peninsula (AQAP) before 1 April 2012?	0.04	0.28	0.24	192	26	0

been provided:

$\hat{p}_G$ = Our aggregate estimate based on the forecasts made within the first three days. The bias term, a, was estimated based on $\hat{a}_{MLE}$.

$\bar{p}$ = Sample average of the forecasts made within the first three days.

s_p = Sample standard deviation of the forecasts made within the first three days.

N = Number of forecasts made within the first three days.

T = Number of days that the problem was open.

Z = Indicator of whether the event happened ($Z = 1$) or not ($Z = 0$).

Even though this paper does not focus on dynamic data, we report the time-frame of each problem, because this is somewhat indicative of the uncertainty and difficulty of the problem (see Table A.1).

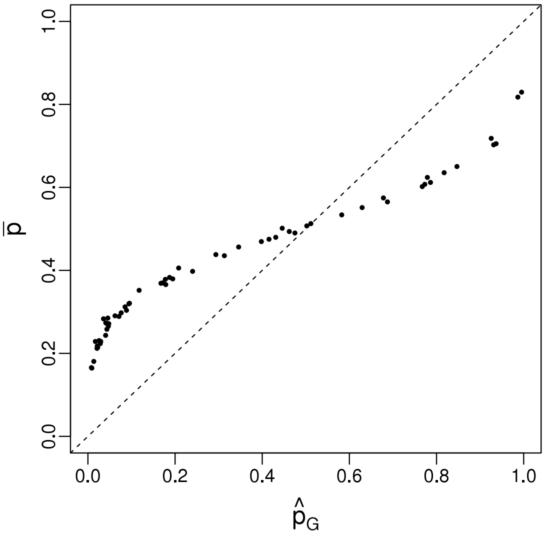

Fig. 7. A summarizing comparison of the aggregators $\hat{p}_G(\hat{a}_{MLE})$ and $\bar{p}$.

Fig. 7 summarizes the data by giving a scatterplot of $\hat{p}_G(\hat{a}_{MLE})$ against $\bar{p}$. Note that the points are above (below) the 45° dashed line when $\hat{p}_G(\hat{a}_{MLE})$ is less (more) than 0.5. This implies that $\hat{p}_G(\hat{a}_{MLE})$ is a much sharper aggregator than $\bar{p}$.

References

Allard, D., Comunian, A., & Renard, P. (2012). Probability aggregation methods in geoscience. *Mathematical Geosciences*, *44*, 545–581.

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., et al. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic Publishers.

Bacharach, M. (1972). *Scientific disagreement*. Unpublished Manuscript.

Baron, J., Ungar, L. H., Mellers, B. A., & Tetlock, P. E. (2013). Two reasons to make aggregated probability forecasts more extreme. (A copy can be requested by emailing Lyle Ungar at ungar@cis.upenn.edu.) (Submitted for publication).

Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, *10*, 1137–1148.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Buja, A., Stuetzle, W., & Shen, Y. (2005). *Loss functions for binary class probability estimation and classification: structure and applications*. Manuscript available at www-stat.wharton.upenn.edu/~buja.

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

Di Bacco, M., Frederic, P., & Lad, F. (2003). *Learning from the probability assertions of experts. Research report*. Manuscript available at http://www.math.canterbury.ac.nz/research/ucdms2003n6.pdf.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: the role of error in judgment processes. *Psychological Review*, *66*, 519–527.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*, 195–200.

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: a critique and an annotated bibliography. *Statistical Science*, *1*, 114–135.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *69*, 243–268.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *14*, 107–114.

Graefe, A., Armstrong, J. S., Jones, R. J., Jr., & Cuzán, A. G. (2014). Combining forecasts: an application to elections. *International Journal of Forecasting*, *30*, 43–54.

McKenzie, C. R., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: what does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*, 179–191.

Pal, S. (2009). *On a conjectured sharpness principle for probabilistic forecasting with calibration*. ArXiv Preprint arXiv:0902.0342.

Polyakova, E. I., & Journel, A. G. (2007). The nu expression for probabilistic data integration. *Mathematical Geology*, *39*, 715–733.

Ranjan, R. (2009). *Combining and evaluating probabilistic forecasts*. Ph.D. Thesis, University of Washington.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *72*, 71–91.

Wallsten, T. S., Budescu, D. V., & Erev, I. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*, 243–268.

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *18*, 1–18.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1–14.

Ville A. Satopää received his B.A. in mathematics and computer science from Williams College in 2011. He is currently a second year Ph.D. student in statistics at the Wharton School of the University of Pennsylvania. He is interested in Bayesian analysis, dynamic models, and non-parametric approaches.

Jonathan Baron is a professor of psychology at the University of Pennsylvania, with interests in moral judgment and, more generally, the field of judgment and decision making and its implications for public policy. He is the editor of the journal *Judgment and Decision Making*.

Dean P. Foster received his B.S. from the University of Maryland in 1980 and his Ph.D. in Mathematics from MD in 1988. He is currently the Marie and Joseph Melone Professor of Statistics at the Wharton School of the University of Pennsylvania. His current research interests are machine learning, stepwise regression and computational linguistics.

Barbara A. Mellers is George I. Heyman University Professor, University of Pennsylvania. She has appointments in the Department of Psychology and the Wharton Marketing Department. She serves as Associate and Consulting Editors on numerous academic journals and has received several grants from NSF. Her research focuses on models of human judgment and decision making. She studies how and why people deviate from principles of rationality and how they can learn to do better. She is currently interested in how people make forecasts and how those forecasts can be improved.

Philip E. Tetlock is the Annenberg University Professor at the University of Pennsylvania, with cross appointments in the Department of Psychology and the Wharton School. He is also a principal investigator for the Intelligence Advanced Research Projects Activities contract that funded the collection of the forecasting data reported in this article. His research focuses on cognitive and motivational biases in human judgment—and methods of debiasing judgment.

Lyle H. Ungar is an Associate Professor of Computer and Information Science (CIS) at the University of Pennsylvania. He also holds appointments in several other departments in the Engineering, Medicine, and Business Schools at Penn, and serves as the Associate Director of the Penn Center for BioInformatics (PCBI). Dr. Ungar received a B.S. from Stanford University and a Ph.D. from M.I.T. He has published over 100 articles, and is co-inventor on eight patents. His research areas include data and text mining, bioinformatics, machine learning, and auction design with a current focus on statistical natural language processing.