Submitted to the Annals of Applied Statistics

PROBABILITY AGGREGATION: DYNAMIC HIERARCHICAL MODELING OF SPARSE EXPERT BELIEFS

BY VILLE A. SATOPÄÄ, SHANE T. JENSEN, LYLE H. UNGAR, BARBARA A. MELLERS, AND PHIL E. TETLOCK

Department of Statistics, The Wharton School of the University of Pennsylvania

E-mail: satopaa@wharton.upenn.edu; stjensen@wharton.upenn.edu

Department of Computer and Information Science, University of Pennsylvania

E-mail: ungar@cis.upenn.edu

Department of Psychology, University of Pennsylvania

E-mail: mellers@wharton.upenn.edu; tetlock@wharton.upenn.edu

Most subjective probability aggregation procedures use a single probability judgement from each expert, even though it is common for experts studying real problems to update their probability estimates over time. This paper makes an advance towards unexplored areas of probability aggregation by considering a dynamic context in which experts are allowed to update their beliefs at random intervals. The updates are assumed to occur very infrequently, resulting into a highly sparse dataset that cannot be modeled by standard time-series procedures. In response to the lack of appropriate methodology, this paper presents a hierarchical model that takes into account the expert's level of self-reported expertise and produces aggregate probabilities that are sharp and well-calibrated both in- and out-of-sample. The model is demonstrated on a real-world dataset that includes over 2,300 experts making multiple probability forecasts on different subsets of 166 international political events.

Keywords and phrases: Probability Aggregation, Dynamic Linear Model, Hierarchical Modeling, Expert Forecast, Subjective Probability, Bias Estimation, Calibration, Time Series

1. Introduction. Individual experts can differ radically from one another in their abilities to assess probabilities of future events. Their probability assessments are often evaluated and compared in regards to *calibration* that measures how closely the frequency of occurrence of the events agree with the given probabilities. For instance, the proportion of occurred events is 60% for all those events for which a well-calibrated expert assessed a probability of 0.60. Even though several experiments have been conducted to show that experts are generally poorly calibrated (see, e.g., Cooke (1991); Shlyakhter et al. (1994)), relative differences can occur among different types of experts. In particular, Wright et al. (1994) argue that a higher level of self-reported expertise can be associated with better calibration.

Calibration by itself, however, is not sufficient for a useful probability assessment. To see this, consider a relatively stationary process, such as rain on different days in some geographic region, whose empirical frequency of occurrence in the last 10 years is 45%. In this setting an expert can assess a constant probability of 0.45 and be well-calibrated. This assessment, however, can be made without any subject-matter expertise or special training. For this reason the long-term frequency is often considered as the baseline probability – a naive assessment that provides the decision-maker very little extra information. Therefore a consulting expert should aim to make probability assessments that are as far from the baseline as possible. The extent to which his probabilities differ from the baseline is measured with an important attribute called *sharpness* (Gneiting et al. (2008); Winkler and Jose (2008)). If the expert is both sharp and well-calibrated, he is able to forecast the behavior of the process with high certainty and accuracy. From the decision-maker's perspective his advice is highly useful as it involves much information and very little risk. Therefore, to summarize, useful probability estimation should aim to maximize sharpness subject to calibration (see, e.g., Raftery et al. (2005); Murphy and Winkler (1987)). This is a well-defined goal that has led to a wide range of novel and insightful observations in probability forecasting.

One such observation is related to the aggregation of multiple probabilities. There is strong empirical evidence that bringing together the strengths of different experts by combining their

probability forecasts into a single consensus, known as the *crowd belief*, results in better predictive performance. Being motivated by the long list of applications of probability forecasts, including medical diagnosis (Wilson et al. (1998); Pepe (2003), political and socio-economic foresight (Tetlock (2005)), and meteorology (Sanders (1963); Vislocky and Fritsch (1995); Baars and Mass (2005)), researchers have proposed many different approaches to combining probability forecasts (see, e.g., Ranjan and Gneiting (2010); Satopää et al. (2013) for some recent studies, and Genest and Zidek (1986); Wallsten, Budescu and Erev (1997); Clemen and Winkler (2007); Primo et al. (2009) for a comprehensive overview). The general focus, however, has been on developing one-time aggregation procedures that consult the expert's advice only once before the event resolves.

Consequently, many areas of probability aggregation still remain rather unexplored. For instance, consider an investor aiming to assess whether a stock index will finish trading above some threshold on a given date. In order to maximize his overall predictive accuracy, he may consult a group of experts repeatedly over a period of time and adjust his estimate of the aggregate probability accordingly. Since the experts are allowed to update their probability assessments, the aggregation should be performed by taking into account the temporal correlation in their advice. Many standard time-series procedures could be used to perform this aggregation as long as the experts update their advice consistently on a daily basis.

This paper adds another layer of complexity by assuming a heterogeneous set of experts, most of whom only make one or two probability assessments over the hundred or so days before the event resolves. This means that the decision-maker faces a different group of experts every day, with only a few experts returning later on for a second round of advice. The problem at hand is therefore strikingly different from many time-series estimation problems, where one has an observation at every time point – or almost every time point. As a result, standard time-series procedures like ARIMA (see, e.g., Mills (1991)) are not directly applicable to our setting. In response to the lack of appropriate modeling procedures, this paper introduces a time-series model that incorporates self-reported expertise and captures a sharp and well-calibrated estimate of the crowd belief. The

Statistic	Min.	Q_1	Median	Mean	Q_3	Max.
# of Days a Problem is Active	4	35.6	72.0	106.3	145.20	418
# of Experts per Problem	212	543.2	693.5	783.7	983.2	1690
# Forecasts given by each Expert on a Problem	1	1.0	1.0	1.8	2.0	131
# Problems participated by an Expert	1	14.0	36.0	55.0	90.0	166

TABLE 1
Five-number summaries of our real-world data.

model is highly interpretable and can be used for

- the analysis of group-level under- and overconfidence across different levels of self-reported expertise,
- accurate probability forecasts, and
- many question-specific quantities that have easy interpretations and can be used to gain novel insight in the social sciences.

The paper begins with a description of our probability forecasting data that were collected by asking over 2,300 experts to give probability forecasts and to self-assess their level of expertise on a subset of 166 geopolitical binary events. After summarizing the dataset, the paper introduces a dynamic hierarchical model for capturing the crowd belief. The model is estimated in a two-step procedure: first, a sampling step produces constrained parameter estimates via Gibbs sampling (see Geman and Geman (1984) for the original introduction of Gibbs sampling); second, a calibration step transforms these estimates to their unconstrained equivalents via a one-dimension optimization procedure. An extension of this model to polychotomous outcomes is briefly discussed before model evaluation. The first evaluation section uses synthetic data to study the accuracy to which the two-step procedure is able to estimate parameter values. The second evaluation section illustrates different aspects of the model by applying it to our real-world probability forecasting data. The paper concludes with a discussion on future research directions and model limitations.

2. Geopolitical Forecasting Data. The data collection began with a recruitment of 2,365 experts ranging from graduate students to political science faculty and practitioners. The recruiting

Expertise Level	1	2	3	4	5
Frequency (%)	25.3	30.7	33.6	8.2	2.1

TABLE 2

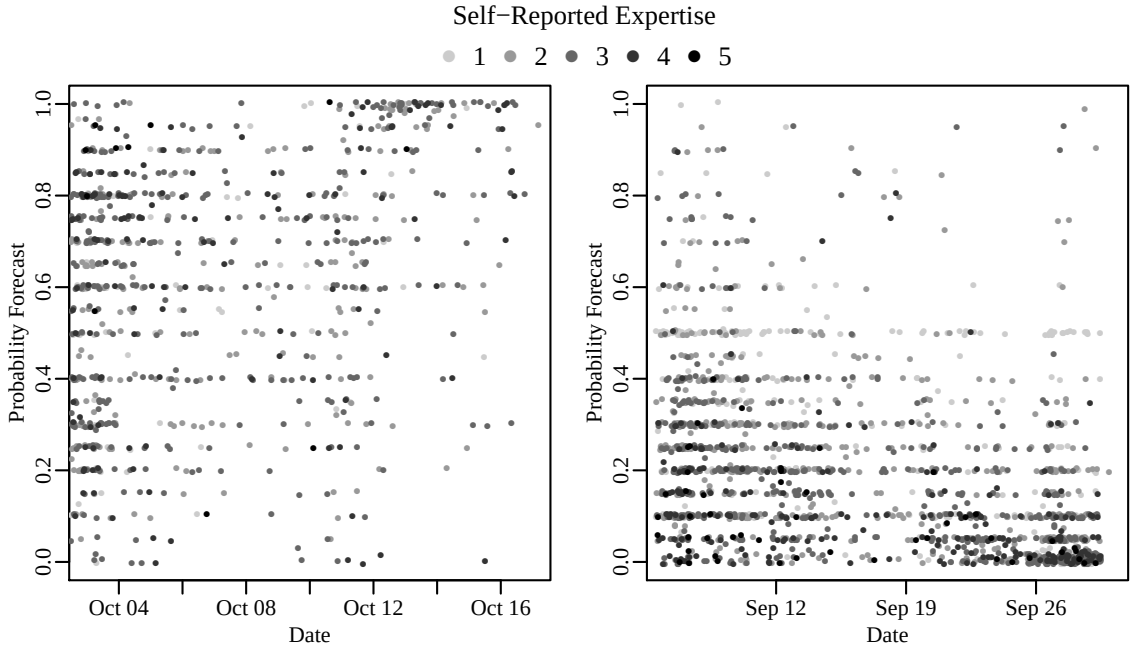
Frequencies of the self-reported expertise (1 = Not At All Expert and 5 = Extremely Expert) levels across all the 166 questions in our real-world data.

was made from professional societies, research centers, alumni associations, science bloggers, and word of mouth. Requirements included at least a Bachelor's degree and 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. The experts were asked to give probability forecasts (to the second decimal point) and to self-assess their level of expertise (on a 1-to-5 scale with 1 = Not At All Expert and 5 = Extremely Expert) on a number of 166 geopolitical binary events taking place between September 29, 2011 and May 8, 2013. Each question was active for a period of time during which the participating experts were allowed to update their forecasts as frequently as they liked without being penalized. The experts knew that their probability estimates would be assessed for accuracy using Brier scores[1]. This incentivized them to report their true beliefs instead of attempting to game the system (Winkler and Murphy (1968)). In addition to receiving $150 for meeting minimum participation requirements that did not depend on prediction accuracy, the experts received status rewards for their performance via leader-boards displaying Brier scores for the top 20 experts. Since a typical expert participated only in a small subset of the 166 questions, the experts are considered indistinguishable conditional on the level of self-reported expertise.

Overall, the data is quite sparse. As mentioned before, the experts were allowed to update their forecasts as long as the question remained open. Updating, however, was done on a very infrequent basis. An expert gave, on average, only 0.016 forecasts per day. Therefore a typical expert made a forecast about every 62 days, resulting to an average response rate of around 13.5 forecasts per

[1]The Brier score is the squared distance between the probability forecast and the event indicator that equals 1.0 or 0.0 depending on whether the event happened or not, respectively. See Brier (1950) for the original introduction.

day from a very large and diverse group of experts. Tables 1 and 2 provide more relevant summary statistics on the data. Notice that the distribution of the self-reported expertise is skewed to the right and that some questions remained active longer than others. For more details on the data and its collection see Ungar et al. (2012).



(a) *Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?* (b) *Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?*

FIG 1. *Scatterplots of the probability forecasts given for two questions in our dataset. The shadings represents the self-reported expertise of the expert who provided the probability forecast.*

To illustrate the nature of the data with some concrete examples, Figures 1(a) and 1(b) show scatterplots of the probability forecasts given for (a) *Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?*, and (b) *Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?*. The points have been jittered slightly to make overlaps visible. The darkness of the points is positively associated with the self-reported expertise. Given that the European bailout fund was ratified before November 1, 2011 and that

the Nikkei 225 index finished trading at around 8,700 on September 30, 2011, the general trend of the probability forecasts tends to converge towards the correct answers. The individual experts, however, disagree rather strongly, with the disagreement persisting even near the closing dates of the questions.

3. Model. Let $p_{i,t,k} \in (0,1)$ be the probability forecast given by the ith expert at time t for the kth question, where $i = 1, \ldots, I_k$, $t = 1, \ldots, T_k$, and $k = 1, \ldots, K$. Denote the logit-probabilities with

$$Y_{i,t,k} = \text{logit}(p_{i,t,k}) = \log\left(\frac{p_{i,t,k}}{1 - p_{i,t,k}}\right) \in \mathbb{R}$$

and collect the logit-probability forecasts given for question k at time t into a vector $\boldsymbol{Y}_{t,k} = [Y_{1,t,k}\ Y_{2,t,k}\ \ldots\ Y_{I_k,t,k}]^T$. Partition the experts into J groups based on some individual feature, such as self-reported expertise, with each group sharing a common multiplicative bias term $b_j \in \mathbb{R}$ for $j = 1, \ldots, J$. Collect these bias terms into a bias vector $\boldsymbol{b} = [b_1\ b_2\ \ldots\ b_J]^T$. Let $\boldsymbol{M}_k$ be a $I_k \times J$ matrix denoting the group-memberships of the experts in question k; that is, if the ith expert participating in the kth question belongs to the jth group, then the ith row of $\boldsymbol{M}_k$ is the jth standard basis vector $\boldsymbol{e}_j$. The bias vector $\boldsymbol{b}$ does not include a subindex because it is considered shared among all the K questions. To secure model identifiability, it is sufficient to share only one of the elements of $\boldsymbol{b}$ among the questions. This element defines a baseline under which it is possible to estimate the remaining $J - 1$ bias terms separately within each of the questions. In this paper, however, the entire vector $\boldsymbol{b}$ is shared because some of the questions in our real-world data set involve very few experts with the highest level of self-reported expertise. Under this notation, the model for the kth question can be expressed as

$$\boldsymbol{Y}_{t,k} = \boldsymbol{M}_k \boldsymbol{b} X_{t,k} + \boldsymbol{v}_{t,k} \tag{1}$$

$$X_{t,k} = \gamma_k X_{t-1,k} + w_{t,k} \tag{2}$$

$$X_{0,k} \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

where Equation (1) denotes the observed process and Equation (2) shows the hidden process that is driven by the constant $\gamma_k \in \mathbb{R}$. The error terms are independent and identically distributed normal random variables with mean zero

$$\begin{aligned} \boldsymbol{v}_{t,k}|\sigma_k^2 &\overset{i.i.d.}{\sim} \mathcal{N}_{I_k}(\mathbf{0}, \sigma_k^2 \boldsymbol{I}_{I_k}) \\ w_{t,k}|\tau_k^2 &\overset{i.i.d.}{\sim} \mathcal{N}(0, \tau_k^2), \end{aligned}$$

and $(\mu_0, \sigma_0^2) \in (\mathbb{R}, \mathbb{R}^+)$ are hyper-parameters chosen *a priori*. The hidden state $X_{t,k}$ represents the sharp and well-calibrated logit-probability for the kth event given the information available up to and including time t. To make this more specific, let $Z_k \in \{0, 1\}$ indicate whether the event associated with the kth question happened $(Z_k = 1)$ or did not happen $(Z_k = 0)$. If $\{\mathcal{F}_{t,k}\}_{t=1}^{T_k}$ is a filtration representing the information available up to and including a given time point, then $\mathbb{E}[Z_k|\mathcal{F}_{t,k}] = \mathbb{P}(Z_k = 1|\mathcal{F}_{t,k}) = \text{logit}^{-1}(X_{t,k})$. These probabilities are made estimable by assuming that the observed process is associated with the filtration $\{\mathcal{F}_{t,k}\}_{t=1}^{T_k}$ and hence can be used as a proxy for the amount of information available at any given time point.

To give intuitive interpretations of the other model parameters, notice that the error term in the observed process has mean zero. Therefore the experts are assumed to be, on average, a multiplicative constant $\boldsymbol{b}$ away from the calibrated crowd belief. An individual element of $\boldsymbol{b}$ can be interpreted as a group-specific *systematic bias* that labels the group either as over-confident $(b_j \in (1, \infty))$ or as under-confident $(b_j \in (0, 1))$. Unfortunately, due to the high sparsity of our data, estimating a bias term and analyzing the bias separately for each expert is not possible. See Section 7.4 for an analysis and discussion on the bias terms. Any other deviation from the calibrated crowd belief is considered *random noise*. This noise is measured in terms of σ_k^2 and can be assumed to be caused by momentary over-optimism (or pessimism), false beliefs, or other misconceptions.

Similarly to the observed process, the hidden process has a systematic and a random component. The *random fluctuations* are measured in terms of τ_k^2 and can be assumed to represent changes or shocks to the underlying circumstances that ultimately decide the outcome of the event. The

systematic component γ_k, on the other hand, allows the model to incorporate a constant signal stream that drifts the hidden process to infinity (when $\gamma_k \in (1, \infty)$) or zero (when $\gamma_k \in (0, 1)$). If the uncertainty in the question diminishes as the current time point t approaches T_k, the hidden process drifts to infinity. Alternatively, the hidden process can drift to zero in which case any available information about the target event does not improve predictive accuracy. Since each of the K questions in our dataset was resolved within a pre-specified timeframe, γ_k is expected to fall within the interval $(1, \infty)$ for all $k = 1, \ldots, K$.

4. Model Estimation. The main challenge is to capture a well-calibrated estimate of the hidden process without sacrificing the interpretability of our model. This section introduces a two-step procedure, called *Sample-Then-Calibrate* (STC), that achieves this goal in a flexible and efficient manner: the first step estimates the model parameters under a constraint (*Sampling Step*), and the second step performs a one-dimension optimization procedure to transform the constrained estimates into their unconstrained counterparts (*Calibration Step*).

4.1. *Sampling Step.* Since $\left(a\boldsymbol{b}, X_{t,k}/a, a^2\tau_k^2\right) \neq \left(\boldsymbol{b}, X_{t,k}, \tau_k^2\right)$ for any $a > 0$ yield the same likelihood for $\boldsymbol{Y}_{t,k}$, the model as described by Equations (1) and (2) is not identifiable. As a result, the parameter estimates tend to drift during the sampling process. A well-known solution is to choose one of the elements of $\boldsymbol{b}$, say b_3, as the reference point and fix $b_3 = 1$. Denote the constrained version of the model by

$$\begin{aligned}
\boldsymbol{Y}_{t,k} &= \boldsymbol{M}_k \boldsymbol{b}(1) X_{t,k}(1) + \boldsymbol{v}_{t,k} \\
X_{t,k}(1) &= \gamma_k(1) X_{t-1,k}(1) + w_{t,k} \\
\boldsymbol{v}_{t,k} | \sigma_k^2(1) &\overset{i.i.d.}{\sim} \mathcal{N}_{I_k}(\boldsymbol{0}, \sigma_k^2(1) \boldsymbol{I}_{I_k}) \\
w_{t,k} | \tau_k^2(1) &\overset{i.i.d.}{\sim} \mathcal{N}\left(0, \tau_k^2(1)\right),
\end{aligned}$$

where the trailing input parameter emphasizes the constraint $b_3 = 1$. Since this version is identifiable, estimates of the model parameters can be obtained. Denote the estimates by placing a hat on

the parameter symbol. For instance, $\hat{\boldsymbol{b}}(1)$ and $\hat{X}_{t,k}(1)$ represent the estimates of $\boldsymbol{b}(1)$ and $X_{t,k}(1)$, respectively. These estimates are found via Gibbs sampling that only makes use of standard distributions. See Appendix A for the technical details of the sampling step, and, e.g., Gelman et al. (2003) for a discussion on the general principles of Gibbs sampling.

4.2. *Calibration Step.* Since the model parameters can be estimated by fixing b_3 to any constant, the next step is to search for the constant that gives an optimally sharp and calibrated estimate of the hidden process. This section introduces an efficient procedure that finds the optimal constraint without requiring any additional runs of the sampling step. First, assume that parameter estimates $\hat{\boldsymbol{b}}(1)$ and $\hat{X}_{t,k}(1)$ have already been obtained via the constrained sampling step described in Section 4.1. Since for any $\beta \in \mathbb{R}/\{0\}$,

$$\begin{aligned} \boldsymbol{Y}_{t,k} &= \boldsymbol{M}_k \boldsymbol{b}(1) X_{t,k}(1) + \boldsymbol{v}_{t,k} \\ &= \boldsymbol{M}_k \left(\boldsymbol{b}(1)\beta\right) \left(X_{t,k}(1)/\beta\right) + \boldsymbol{v}_{t,k} \\ &= \boldsymbol{M}_k \boldsymbol{b}(\beta) X_{t,k}(\beta) + \boldsymbol{v}_{t,k}, \end{aligned}$$

the parameter values under $b_3 = \beta$ can be obtained from $\boldsymbol{b}(\beta) = \boldsymbol{b}(1)\beta$ and $X_{t,k}(\beta) = X_{t,k}(1)/\beta$. This means that $X_{t,k} = X_{t,k}(1)/\beta$ when β is equal to the true value of b_3. Since the hidden process $X_{t,k}$ is assumed to be sharp and well-calibrated, b_3 can be estimated with the value of β that simultaneously maximizes the sharpness and calibration of $\hat{X}_{t,k}(1)/\beta$. A natural criterion for this maximization is given by the class of *proper scoring rules* that combine sharpness and calibration (Gneiting et al. (2008); Buja, Stuetzle and Shen (2005)). Due to the possibility of *complete separation* in any one question (see, e.g., Gelman et al. (2008)), the maximization must be performed over multiple questions. Therefore,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}/\{0\}}{\arg\max} \sum_{k=1}^{K} \sum_{t=1}^{T_k} S\left(Z_k, \hat{X}_{k,t}(1)/\beta\right) \tag{3}$$

where $Z_k \in \{0, 1\}$ indicate whether the event associated with the kth question happened ($Z_k = 1$) or did not happen ($Z_k = 0$). The function S is a strictly proper scoring rule such as the negative

Brier score (Brier (1950))

$$S_{BRI}(Z, X) \quad = \quad -(Z - \text{logit}^{-1}(X))^2$$

or the logarithmic score (Good (1952))

$$S_{LOG}(Z, X) \quad = \quad Z \log\left(\text{logit}^{-1}(X)\right) + (1 - Z) \log\left(1 - \text{logit}^{-1}(X)\right)$$

Since it is not clear which rule should be used for predicting geopolitical events, the *Sample-Then-Calibrate* procedure is evaluated separately under both rules in Sections 6 and 7. Once $\hat{\beta}$ has been computed, estimates of the unconstrained model parameters are given by

$$\begin{aligned}
\hat{X}_{t,k} &= \hat{X}_{k,t}(1)/\hat{\beta} \\
\hat{\boldsymbol{b}}_{t,k} &= \hat{\boldsymbol{b}}(1)\hat{\beta} \\
\hat{\tau}_k^2 &= \hat{\tau}_k^2(1)\hat{\beta}^2 \\
\hat{\sigma}_k^2 &= \hat{\sigma}_k^2(1) \\
\hat{\gamma}_k &= \hat{\gamma}_k(1)
\end{aligned}$$

Notice that estimates of σ_k^2 and γ_k are not affected by the constraint. Therefore their constrained and unconstrained versions are the same.

4.3. *Discussion.* If the class labels in the data are balanced with respect to the time points, the calibration step under the logarithmic scoring rule is approximately equivalent to *Platt calibration*, which has been shown to yield good calibration under various modeling scenarios (see, e.g., Platt et al. (1999); Niculescu-Mizil and Caruana (2005)). To see this, recall that the Platt calibrated logit-probabilities are given by $\hat{A} + \hat{B}\hat{X}_{t,k}(1)$, where

$$\left(\hat{A}, \hat{B}\right) = \underset{A,B \in \mathbb{R}}{\arg\max} \sum_{k=1}^{K} \sum_{t=1}^{T_k} S_{LOG}(Z_k, A + B\hat{X}_{t,k}(1)) \tag{4}$$

This is equivalent to fitting a logistic regression model with Z_k as the response and $\hat{X}_{t,k}(1)$ as the explanatory variable. To understand the behavior of the coefficients A and B, express the logistic regression as a linear regression model

$$\text{logit}(\mathbb{P}(Z_k = 1|\hat{X}_{t,k})) = A + B\hat{X}_{t,k} + e_{t,k}$$

with $e_{t,k} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. If the data are balanced with respect to the time points, then exactly half of the summands in Equation (4) have $Z_k = 1$ and the average response logit-probability $\text{logit}(\mathbb{P}(Z_k = 1|\hat{X}_{t,k}))$ is expected to be close to zero. Since the values of $\hat{X}_{t,k}$ are estimated logit-probabilities of the same K events across different time points, their overall average is also expected to be around zero. Therefore both the response and explanatory variables are approximately centered. This means that the intercept term A is near zero reducing Platt calibration to Equation (3) under the logarithmic scoring rule. If the data are not balanced, Platt calibration can be easily incorporated into our model via an additional intercept parameter. This, however, reduces the interpretability of our model. Fortunately, compromising interpretability is rarely necessary because it is often possible to use the data in a well-balanced form. One procedure to attain this is described in the beginning of Section 7.

5. Extension: Polychotomous Outcomes. If the future event can take upon $M > 2$ possible outcomes, the hidden state $X_{t,k}$ must be extended to a vector of size $M - 1$. One of the outcomes, e.g., the Mth one, is chosen as the base-case. This gives us a total of $M - 1$ observed processes and one hidden process

$$\begin{aligned}
\boldsymbol{Y}_{1,t,k} &= \boldsymbol{M}_k \boldsymbol{b}_1 X_{1,t,k} + \boldsymbol{v}_{1,t,k} \\
\boldsymbol{Y}_{2,t,k} &= \boldsymbol{M}_k \boldsymbol{b}_2 X_{2,t,k} + \boldsymbol{v}_{2,t,k} \\
&\vdots \\
\boldsymbol{Y}_{M-1,t,k} &= \boldsymbol{M}_k \boldsymbol{b}_{M-1} X_{M-1,t,k} + \boldsymbol{v}_{M-1,t,k} \\
\boldsymbol{X}_{t,k} &= \boldsymbol{\gamma}_k^T \boldsymbol{I}_{M-1} \boldsymbol{X}_{t-1,k} + \boldsymbol{w}_{t,k}
\end{aligned}$$

where $\boldsymbol{X}_{t,k}$ is a $(M-1)\times 1$ matrix of calibrated logit-probabilities. Notice that each process has a separate bias vector. The error terms are independent and identically distributed multivariate normal random variables

$$\begin{aligned}\boldsymbol{v}_{m,t,k}|\sigma^2_{j,k} &\overset{i.i.d.}{\sim} \mathcal{N}_{I_k}\left(\mathbf{0}, \sigma^2_{j,k}\boldsymbol{I}_{I_k}\right)\\ \boldsymbol{w}_{t,k}|\tau^2_k &\overset{i.i.d.}{\sim} \mathcal{N}_{M-1}\left(\mathbf{0}, \tau^2_k\boldsymbol{I}_{M-1}\right)\end{aligned}$$

On the left-hand side of observed process

$$\begin{aligned}\boldsymbol{Y}_{m,t,k} &= \begin{bmatrix}Y_{m,1,t,k} & Y_{m,2,t,k} & \dots & Y_{m,I_k,t,k}\end{bmatrix}^T\\ &= \begin{bmatrix}\log\left(\frac{p_{m,1,t,k}}{p_{M,1,t,k}}\right) & \log\left(\frac{p_{m,2,t,k}}{p_{M,2,t,k}}\right) & \dots & \log\left(\frac{p_{m,I_k,t}}{p_{M,I_k,t,k}}\right)\end{bmatrix}^T\end{aligned}$$

is a $I_k\times 1$ matrix of the expert logit-probabilities for the mth outcome of the kth question at time t, and on the left-hand side of the hidden process

$$\boldsymbol{X}_{t,k} = \begin{bmatrix}X_{1,t,k} & X_{2,t,k} & \dots & X_{M-1,t,k}\end{bmatrix}^T$$

is the matrix of sharp and calibrated logit-probabilities at time t. Choosing one of the outcomes as the base-case, ensures that the probabilities will sum to one at any given time point. Since this multinomial extension is equivalent to having $M-1$ independent binary-outcome models (see Section 3), the estimation can be done separately for each outcome. Therefore, even though this paper focuses on modeling binary events, all the properties and discussion generalize to the multi-outcome case.

6. Synthetic Data Results. The synthetic data is not generated directly from the model for two reasons: (i) showing good performance on a dataset directly generated from the model assumptions is hardly any news, and (ii) generating data from the dynamic model description does not produce well-calibrated hidden states. The latter is important for our interpretation of the hidden process as a sharp and well-calibrated version of the crowd belief. To ensure that the hidden process for

question k is well-calibrated, generate a path of the standard Brownian motion until time T_k. If $Z_{t,k}$ denotes the value of the path at time t, then

$$\begin{aligned} Z_k &= \mathbb{1}\left(Z_{T_k,k} > 0\right) \\ X_{t,k} &= \text{logit}\left[\Phi\left(\frac{Z_{t,k}}{\sqrt{T_k - t}}\right)\right] \end{aligned}$$

gives a sequence of T_k calibrated logit-probabilities for the event $Z_k = 1$.

We use this procedure to generate the hidden processes for K questions with $T_k = 100$ for $k = 1, \ldots, K$. Each question involves 50 experts allocated evenly among the five expertise groups. The expert logit-probabilities $Y_{i,t,k}$ are generated from Equation (1) by applying bias and noise to the hidden process $X_{t,k}$. Any extreme logit-probabilities are truncated to avoid overly influential observations; that is, any $p_{i,t,k} = \text{logit}^{-1}(Y_{i,t,k})$ less than 0.01 (or greater than 0.99) is truncated to 0.01 (or to 0.99). This is akin to the real-world data setting, where extreme probabilities must be truncated (see Section 7). Therefore, to understand the effect that truncation has on the real-world data results, truncation should be first analyzed and understood under the controlled environment of synthetic data.

Throughout the evaluation, data is generated under different choices of parameter values. More specifically, the evaluation procedure iterates over a three-dimensional grid of parameter values

$$\begin{aligned} \beta &\in \{1/2, 3/4, 1, 4/3, 2/1\} \\ \sigma^2 &\in \{0.5, 1.0, 1.5, 2.0, 2.5\} \\ K &\in \{20, 35, 50, 65, 80\}, \end{aligned}$$

where β generates the bias vector $\boldsymbol{b} = [0.50, 0.75, \ldots, 1.50]^T\beta$. These values were considered to be realistic based on our experience with the real-world probability forecasting data. Each triple of the parameter values, i.e. a grid point, is used 40 times to generate a synthetic dataset. Each dataset is then used to assess how accurately *Sample-Then-Calibrate* is able to estimate the hidden process and bias vector. The accuracy of the estimated hidden process is measured with the average

quadratic loss in the probability space, $\sum_{k=1}^{K}\sum_{t=1}^{T_k}(\text{logit}^{-1}(\hat{X}_{t,k}) - \text{logit}^{-1}(X_{t,k}))^2/\sum_{k=1}^{K} T_k$, and the accuracy of the estimtated bias vector is quantified by the average quadratic loss, $||\boldsymbol{b} - \hat{\boldsymbol{b}}||^2/5$.

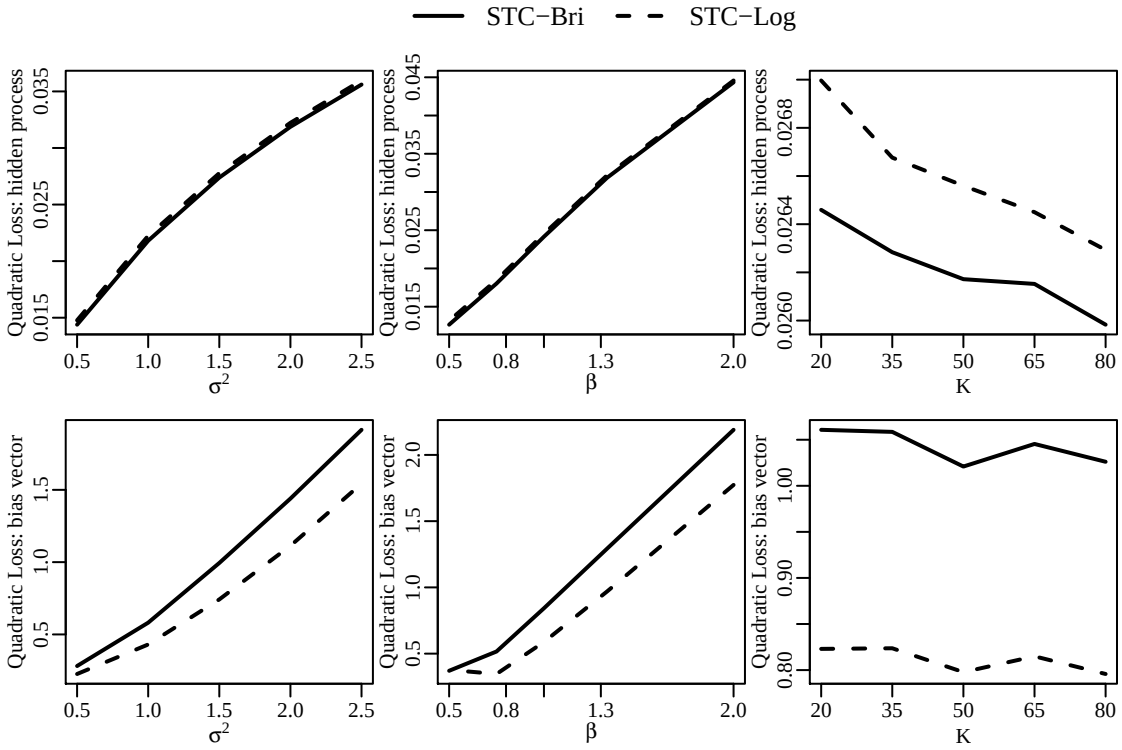


FIG 2. *Comparing Sample-Then-Calibrate that optimizes over the logarithmic score (STC-Log) and Sample-Then-Calibrate that optimizes over the Brier score (STC-Bri) under synthetic data. The top row presents the accuracy to capture the hidden process. The bottom row presents the accuracy to capture the bias vector.*

Figure 2 summarizes the results in a series of plots. The first row shows the marginal effects of the three grid variables on the accuracy to which the *Sample-Then-Calibrate* under the logarithmic loss (STC-Log) and the Brier score (STC-Bri) are able to estimate the hidden process. The second row is similar in nature but measures the accuracy of estimating the bias vector instead. The plots were computed by choosing one grid variable (e.g. β) at a time and averaging over the remaining two variables (e.g. K, and σ^2). Therefore each value in the plots represents an average of a $40 \times$

$5^2 = 1,000$ values.

Based on the first row of plots, STC-Log and STC-Bri estimate the hidden process to a similar accuracy. Even though some relative differences are visible, the loss values are very small in absolute terms, hence suggesting good overall accuracy to capture the hidden process. Estimating the hidden process, however, becomes more difficult as either σ^2 or β gets larger. The latter seems more unintuitive and can be understood better by observing that larger values of β lead to more extreme expert logit-probabilities and hence to more frequent truncation. This causes the estimation accuracy to degrade because a high level of truncation results to a drastic deviation from the model assumptions. Increasing the number of questions improves the estimation accuracy but only very little.

Based on the second row of plots, STC-Log and STC-Bri capture the bias vector rather precisely as long as the true values of β and σ^2 are not very large. This behavior can be again considered as being caused by the truncation of the extreme expert logit-probabilities. Surprisingly, even though STC-Log is more accurate in estimating the bias vector, it is unable to outperform STC-Bri in estimating the hidden process.

7. Geopolitical Data Results. This section discusses results related to the real-world data described in Section 2. The goal is to provide application specific insight by discussing the specific research objectives itemized in Section 1 and also to evaluate the *Sample-Then-Calibrate*-procedure in terms of predictive power and calibration. Before presenting the results, however, we discuss two practical matters that must be taken into account when aggregating real-world probability forecasts.

7.1. *Incoherent and Imbalanced Data.* The first matter regards human experts making probability forecasts of 0.0 or 1.0 even if they are not completely sure of the outcome of the event. For instance, all the 166 questions in our dataset contained both a zero and a one. Transforming such forecasts into the logit-space yields infinities that are highly influential and cause problems in model estimation. To avoid this, Ariely et al. (2000) suggest changing $p = 0.00$ and 1.00 to

$p = 0.02$ and 0.98, respectively. This is similar to *winsorising* that sets the extreme probabilities to a specified percentile of the data (see, e.g., Hastings et al. (1947) for more details on winsorising). Allard, Comunian and Renard (2012), on the other hand, consider only probabilities that fall within a constrained interval, say $[0.001, 0.999]$, and discard the rest. Since this implies ignoring a portion of the data, we have decided to adopt the first approach by changing $p = 0.00$ and 1.00 to $p = 0.01$ and 0.99, respectively.

The second matter is related to the distribution of the class labels in the data. If the set of occurred events is much larger than the set of events that did not occur (or *vice versa*), the dataset is called *imbalanced*. On such data the modeling procedure can end up over-focusing on the larger class, and as a result, give very accurate forecast performance over the larger class at the cost of performing poorly over the smaller class (see, e.g., Chen (2009); Wallace and Dahabreh (2012)). Fortunately, it is often possible to use a well-balanced version of the data. The first step is to find a partition S_0 and S_1 of the question indices $\{1, 2, \ldots, K\}$ such that the equality $\sum_{k \in S_0} T_k = \sum_{k \in S_1} T_k$ is as closely approximated as possible. This is equivalent to an NP-hard problem known in computer science as the *Partition Problem*: determine whether a given set of positive integers can be partitioned into two sets such that the sums of the two sets equal to each other (see, e.g., Karmarkar and Karp (1982); Hayes (2002)). A simple solution is to use a greedy algorithm that iterates through the values of T_k in descending order, assigning each T_k to the subset that currently has the smaller sum (see, e.g. Kellerer, Pferschy and Pisinger (2004); Gent and Walsh (1996) for more details on the Partition Problem). After finding a well-balanced partition, the next step is to assign the class labels such that the labels for the questions in S_x are equal to x for $x = 0$ or 1. Recall from section 4.2 that Z_k represents the event indicator associated with the kth question. To define a balanced set of indicators $\tilde{Z}_k$ for all $k \in S_x$, let

$$\tilde{Z}_k = x$$

$$\tilde{p}_{i,t,k} = \begin{cases} 1 - p_{i,t,k} & \text{if } Z_k = 1 - x \\ p_{i,t,k} & \text{if } Z_k = x, \end{cases}$$

where $i = 1, \ldots, I_k$, and $t = 1, \ldots, T_k$. The resulting set

$$\left\{ \left(\tilde{Z}_k, \{\tilde{p}_{i,t,k} | i = 1, \ldots, I_k, t = 1, \ldots, T_k\} \right) \right\}_{k=1}^{K}$$

is a balanced version of the data. This procedure was used to balance our real-world dataset both in terms of events and time points. The final output splits the events exactly in half ($|S_0| = |S_1| = 83$) such that number of time points in the first and second halves are 8,737 and 8,738, respectively.

7.2. *Out-of-Sample Forecasting.* This section evaluates the out-of-sample predictive performance of *Sample-Then-Calibrate* against several other probability aggregation procedures. The models are allowed to utilize a training set before making predictions on an independent testing set. In order to clarify some of the upcoming notation, let S_{train} and S_{test} be index sets that partition the data into training and testing sets of sizes $|S_{train}| = N_{train}$ and $|S_{test}| = 166 - N_{test}$, respectively. This means that the kth question is in the training set if and only if $k \in S_{train}$. The competing models are as follows.

1. *Simple Dynamic Linear Model (SDLM)*. This is equivalent to the dynamic model from Section 3 but with $\boldsymbol{b} = \mathbf{1}$ and $\beta = 1$. Thus,

$$\begin{aligned} \boldsymbol{Y}_{t,k} &= X_{t,k} + \boldsymbol{v}_{t,k} \\ X_{t,k} &= \gamma_k X_{t-1,k} + w_{t,k}, \end{aligned}$$

   where $X_{t,k}$ is the logit-probability used for prediction. Since this model is not hierarchical, estimates of the hidden process can be obtained directly for the questions in the testing set without fitting the model first on the training set. In order to make predictions, the sampler is run for 500 iterations of which the first 200 are used for burn-in. The remaining 300 iterations

are thinned by discarding every other observation, leaving a final predictive sample of 150 observations.

2. *The Sample-Then-Calibrate procedure both under the Brier (STC-Bri) and the logarithmic score (STC-Log)*. The model is first fit on the training set by running the sampling step for 3,000 iterations of which the first 500 iterations are used for burn-in. After thinning by only keeping every fifth observation, the calibration step is performed for the remaining 500 observations. The out-of-sample prediction is done by running the sampling step for 500 iterations with each consecutive iteration reading in and conditioning on the next value of β and $\boldsymbol{b}$ found during the training period. The first 200 iterations are used for burn-in. The remaining 300 iterations are thinned by discarding every other observation, leaving a final predictive sample of 150 observations.
3. *A fully Bayesian version of STC-Log (BSTC-Log)*. Denote the calibrated logit-probabilities and the event indicators across all K questions with $\boldsymbol{X}(1)$ and $\boldsymbol{Z}$, respectively. The posterior distribution of β conditional on $\boldsymbol{X}(1)$ is given by $p(\beta|\boldsymbol{X}(1), \boldsymbol{Z}) \propto p(\boldsymbol{Z}|\beta, \boldsymbol{X}(1))p(\beta|\boldsymbol{X}(1))$. Recall that the calibration step under S_{LOG} is equivalent to fitting a logistic regression model with Z_k as the response and $\hat{X}_{k,t}(1)$ as the explanatory variable. Therefore the likelihood for the Bayesian version is

$$p(\boldsymbol{Z}|\beta, \boldsymbol{X}(1)) \propto \prod_{k=1}^{K}\prod_{t=1}^{T_k} \text{logit}^{-1}\left(X_{k,t}(1)/\beta\right)^{Z_k}\left(1-\text{logit}^{-1}\left(X_{k,t}(1)/\beta\right)\right)^{1-Z_k} \tag{5}$$

Similarly to Gelman et al. (2003), the prior is chosen to be locally uniform, $p(1/\beta) \propto 1$. Posterior estimates of β can be sampled from Equation (5) using generic sampling algorithms such as the Metropolis algorithm (Metropolis et al. (1953)) or slice sampling (Neal (2003)). Since the sampling procedure conditions on the event indicators, the full conditional distribution of the hidden states is not in a standard form. Therefore the Metropolis algorithm is also used for sampling the hidden states. Predictions are made with the same choices of thinning and burn-in as described under *Sample-Then-Calibrate*.

4. Due to the lack of previous literature on dynamic aggregation of expert probability forecasts, the main competitors are exponentially weighted versions of procedures that have been proposed for static probability aggregation:

   (a) *Exponentially Weighted Moving Average (EWMA)*. If

   $$\bar{p}_{t,k} = \frac{1}{N_{t,k}} \sum_{i=1}^{N_{t,k}} p_{i,t,k},$$

   is the average probability forecast given at time t for the kth question, then the EWMA forecasts for the kth problem are obtained recursively from

   $$\hat{p}_{t,k}(\alpha) = \begin{cases} \bar{p}_{1,k} & \text{for } t = 1 \\ \alpha \bar{p}_{t,k} + (1-\alpha)\hat{p}_{t-1,k}(\alpha) & \text{for } t > 1 \end{cases}$$

   where the input parameter α is learned from the training set by

   $$\hat{\alpha} \quad = \quad \underset{\alpha \in [0,1]}{\arg\min} \sum_{k \in S_{train}} \sum_{t=1}^{T_k} (Z_k - \hat{p}_{t,k}(\alpha))^2$$

   (b) *Exponentially Weighted Moving Logit Aggregator (EWMLA)*. This is a moving version of the aggregator $\hat{p}_G(\boldsymbol{b})$ that was introduced in Satopää et al. (2013). If $\boldsymbol{p}_{t,k}$ is a vector collecting all the probability forecasts made for the kth question at time t, then the EWMLA forecasts are found recursively from

   $$\hat{p}_{t,k}(\alpha, \boldsymbol{b}) = \begin{cases} G_{t,k}(\boldsymbol{b}) & \text{for } t = 1 \\ \alpha G_{t,k}(\boldsymbol{b}) + (1-\alpha)\hat{p}_{t-1,k}(\alpha, \boldsymbol{b}) & \text{for } t > 1 \end{cases}$$

   where

   $$G_{t,k}(\nu) \quad = \quad \left(\prod_{i=1}^{N_{t,k}} \left(\frac{p_{i,t,k}}{1-p_{i,t,k}} \right)^{\frac{\boldsymbol{e}_{i,k}'\boldsymbol{b}}{N_{t,k}}} \right) \Bigg/ \left(1 + \prod_{i=1}^{N_{t,k}} \left(\frac{p_{i,t,k}}{1-p_{i,t,k}} \right)^{\frac{\boldsymbol{e}_{i,k}'\boldsymbol{b}}{N_{t,k}}} \right)$$

   The vector $\boldsymbol{b}$ collects the bias terms of the different expertise groups. Therefore it is equivalent to the bias vector found under *Sample-Then-Calibrate*. The term $\boldsymbol{e}_{i,k}$ is a

vector of length 5 indicating which level of self-reported expertise the ith expert in the kth question belongs to. For instance, if $\boldsymbol{e}_{i,k} = [0, 1, 0, 0, 0]$, then the expert identifies himself with the expertise level two. The tuning parameters $(\alpha, \boldsymbol{b})$ are learned from the training set by

$$(\hat{\alpha}, \hat{\boldsymbol{b}}) \quad = \quad \underset{\boldsymbol{b}\in\mathbb{R}^5, \alpha\in[0,1]}{\arg\min} \sum_{k\in S_{train}} \sum_{t=1}^{T_k} \left(Z_k - \hat{p}_{t,k}(\alpha, \boldsymbol{b})\right)^2$$

(c) *Exponentially Weighted Moving Beta-transformed Aggregator (EWMBA).* The static version of the Beta-transformed aggregator was introduced in Ranjan and Gneiting (2010). A dynamic version can be obtained by replacing $G_{t,k}(\nu)$ in the EWMLA description with

$$H_{\nu,\tau}\left(\bar{p}_{t,k}\right),$$

where $H_{\nu,\tau}$ is the cumulative distribution function of the Beta distribution and $\bar{p}_{t,k}$ is the average probability forecast defined under EWMA. The tuning parameters (α, ν, τ) are learned from the training set by

$$(\hat{\alpha}, \hat{\nu}, \hat{\tau}) \quad = \quad \underset{\nu,\tau>0\ \alpha\in[0,1]}{\arg\min} \sum_{k\in S_{train}} \sum_{t=1}^{T_k} \left(Z_k - \hat{p}_{t,k}(\alpha, \nu, \tau)\right)^2$$

The competing models are evaluated via a 10-fold cross-validation[2] that first partitions the 166 questions into 10 sets. The partition is chosen such that each of the 10 sets has approximately the same number of questions (16 or 17 questions per set in our case) and the same number of time points (between 1760 and 1764 time points per set in our case). The evaluation then iterates 10 times, each time using one of the 10 sets as the testing set and the remaining 9 sets as the training set. Therefore each question is used nine times for training and exactly once for testing. The testing proceeds sequentially one testing question at a time as follows: First, for a question with a time

[2] A 5-fold cross-validation was also performed. The results were, however, very similar to the 10-fold cross-validation and hence not presented in the paper.

horizon of T_k, make a prediction based on the first two days. Compute the Brier score for the aggregate forecast of the second day. Next make a prediction based on the first three days and compute the Brier score for the most recent day, namely, the third day. Repeat this process until the prediction is made on all of the $T_k - 1$ days. This leads to $T_k - 1$ Brier scores per testing question and a total of 17,475 Brier scores across the entire dataset.

Table 3 summarizes different ways to aggregate these scores. The first option, denoted by *Scores by Day*, weighs each question by the number of days the question remained open. This is performed by computing the average of the 17,475 scores. The second option, denoted by *Scores by Problem*, gives each question an equal weight regardless how long the question remained open. This is done by first averaging the scores within a question and then averaging the average scores across all the questions. Both scores can be further broken down into subcategories by considering the length of the questions. The final three columns of Table 3 divide the questions into *Short* questions (30 days or fewer), *Medium* questions (between 31 and 59 days), and *Long* Problems (60 days or more). The number of questions in these subcategories were 36, 32 and 98, respectively. The bolded scores indicate the lowest score in each column. The values in the parenthesis quantify the variability in the scores: Under *Scores by Day* the values give the standard errors of all the scores. Under *Scores by Problem*, on other hand, the values represent the standard errors of the average scores of the different questions.

Overall, STC-Bri and STC-Log achieve the lowest average scores across all columns except *Short*, where they are slightly outperformed by BSTC-Log. BSTC-Log, however, turns out to be overconfident (see Section 7.3). This means that BSTC-Log underestimates the uncertainty in the events and outputs probability forecasts that are typically too near 0.0 or 1.0. As a result, the aggregate forecasts are either very close to the correct answer or very far from it. As can be seen in Table 3, this results into highly variable forecasting performance. The short questions, however, involve very little uncertainty and were generally the easiest to forecast. On such easy questions, overconfidence can pay off frequently enough to compensate for a few large scores arising from

Scores by Day

Model	All	Short	Medium	Long
SDLM	0.100 (0.156)	0.066 (0.116)	0.098 (0.154)	0.102 (0.157)
BSTC-Log	0.097 (0.213)	**0.053** (0.147)	0.100 (0.215)	0.098 (0.215)
STC-Bri	0.096 (0.190)	0.056 (0.134)	0.097 (0.190)	0.098 (0.192)
STC-Log	**0.096** (0.191)	0.056 (0.134)	**0.096** (0.189)	**0.098** (0.193)
EWMBA	0.102 (0.203)	0.060 (0.124)	0.110 (0.201)	0.103 (0.206)
EWMLA	0.102 (0.199)	0.061 (0.130)	0.111 (0.214)	0.103 (0.200)
EWMA	0.111 (0.142)	0.089 (0.100)	0.111 (0.136)	0.112 (0.144)

Scores by Problem

Model	All	Short	Medium	Long
SDLM	0.089 (0.116)	0.064 (0.085)	0.106 (0.141)	0.092 (0.117)
BSTC-Log	0.083 (0.160)	**0.052** (0.103)	0.110 (0.198)	0.085 (0.162)
STC-Bri	0.083 (0.142)	0.055 (0.096)	0.106 (0.174)	0.085 (0.144)
STC-Log	**0.082** (0.142)	0.055 (0.096)	**0.105** (0.174)	**0.085** (0.144)
EWMBA	0.090 (0.156)	0.063 (0.101)	0.118 (0.186)	0.091 (0.161)
EWMLA	0.090 (0.159)	0.064 (0.109)	0.120 (0.200)	0.090 (0.159)
EWMA	0.104 (0.105)	0.092 (0.081)	0.119 (0.125)	0.103 (0.107)

TABLE 3

Brier Scores based on 10-fold cross-validation. Scores by Day weighs a question by the number of days the question remained open. Scores by Problem gives each question an equal weight regardless how long the question remained open. The bolded values indicate the lowest scores in each column. The values in the parenthesis represent standard errors in the scores.

the overconfident and incorrect forecasts.

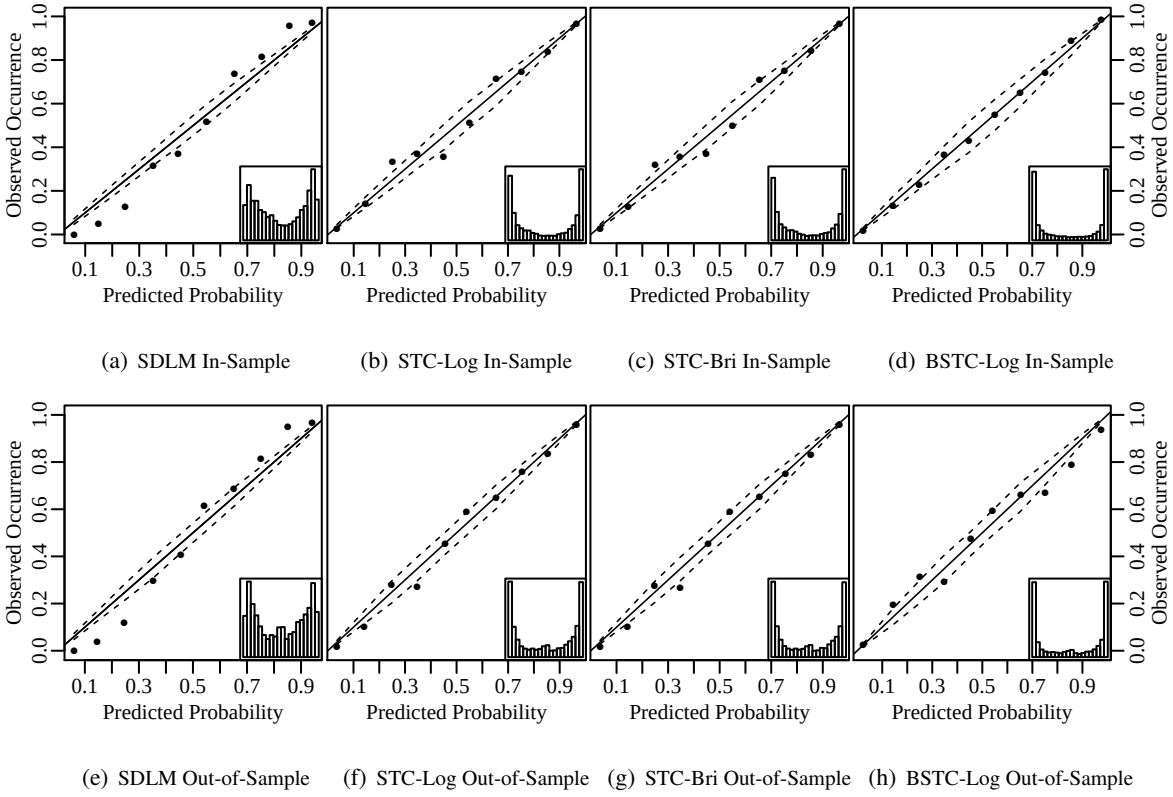
In comparison to BSTC-Log, the baseline SDLM-model lacks sharpness and is highly under-confident (see Section 7.3). This behavior is expected as the experts are under-confident at the group-level (see Section 7.4) and the SDLM-procedure does not use the training set to explicitly calibrate its forecasts. Instead, it merely smooths the forecasts given by the experts. The resulting aggregate forecasts are therefore necessarily conservative, resulting into high average scores with low variability.

Similarly behavior is exhibited by EWMA that performs the worst of all the competing models. The other two exponentially weighted aggregators, namely EWMLA and EWMBA, on other hand, make efficient use of the training set and present good forecasting performance in most columns of Table 3. Recall that EWMBA uses the cumulative distribution function of the Beta distribution to transform the average forecast probability. This transformation depends on two parameters and

is more flexible than the transformation used by EWMLA. On other hand, only EWLMA is given access to the self-reported expertise information. Based on the results, however, it is unable to use this information to outperform EWMBA. Overall, these two aggregators perform very similarly. Unfortunately, they both suffer from over-confidence and are hence unable to outperform the *Sample-Then-Calibrate*-procedure.

7.3. *In- and Out-of-Sample Sharpness and Calibration.* A calibration plot is a simple tool for assessing the sharpness and calibration of a model. The idea is to plot the probability forecasts against the observed empirical frequencies. Therefore any deviation from the diagonal line suggests poor calibration. A model is considered under-confident (or over-confidence) if the points follow an S-shaped (or Ƨ-shaped) trend. In order to assess sharpness of the model, it is common practice to place a histogram of the given forecasts in the corner of the plot. Since the data were balanced, any deviation from the the baseline probability of 0.5 suggests improved sharpness.

The top and bottom rows of Figure 3 present calibration plots for SDLM, STC-Log, STC-Bri, and BSTC-Log under in- and out-of-sample probability estimation, respectively. Each setting is of interest in its own right: Good in-sample calibration is crucial for model interpretability. In particular, if the estimated crowd belief is well-calibrated, then the elements of the bias vector $\boldsymbol{b}$ can be used to study the amount of under- or over-confidence in the different expertise groups. Good out-of-sample calibration and sharpness, on other hand, are necessary properties in predicting future events with high accuracy. To guide our assessment, the dashed bands around the diagonal connect the point-wise, Bonferroni-corrected (Bonferroni (1936)) 95% lower and upper critical values under the null hypothesis of calibration. These have been computed by running the bootstrap technique described in Bröcker and Smith (2007) for 10,000 iterations. The in-sample predictions were obtained by running the models for 10,200 iterations, leading to a final posterior sample of 1,000 observations after thinning and using the first 200 iterations for burn-in. The out-of-sample predictions were given by the 10-fold cross-validation discussed in Section 7.2.



(a) SDLM In-Sample (b) STC-Log In-Sample (c) STC-Bri In-Sample (d) BSTC-Log In-Sample

(e) SDLM Out-of-Sample (f) STC-Log Out-of-Sample (g) STC-Bri Out-of-Sample (h) BSTC-Log Out-of-Sample

FIG 3. *The top and bottom rows show in- and out-of-sample calibration and sharpness, respectively. The models are Simple Dynamic Linear Model (SDLM), the Sample-Then-Calibrate approach that optimizes over the logarithmic score (STC-Log), the Sample-Then-Calibrate approach that optimizes over the Brier score (STC-Bri), and the fully-Bayesian version of the Sample-Then-Calibrate approach that optimizes over the logarithmic score (BSTC-Log).*

Overall, the *Sample-Then-Calibrate*-procedure is sharp and well-calibrated both in- and out-of-sample with only a few points barely falling outside the *point-wise* critical values. Since the calibration does not change drastically from the top to the bottom row, the *Sample-Then-Calibrate*-procedure can be considered to present robustness against over-fitting. This is, however, not the case with BSTC-Log that is well-calibrated in-sample but presents over-confidence out-of-sample. Figures 3(a) and 3(e) serve as baselines by showing the reliability plots for the SDLM-model. Since this model does not perform any explicit calibration, it is not surprising to see most points outside the critical values. The pattern in the deviations suggests drastic under-confidence. Furthermore,

the inset histogram reveals drastic lack of sharpness. Therefore the *Sample-Then-Calibrate*-model can be viewed as a well-performing compromise between SDLM and BSTC-Log that avoids over-confidence without being too conservative.

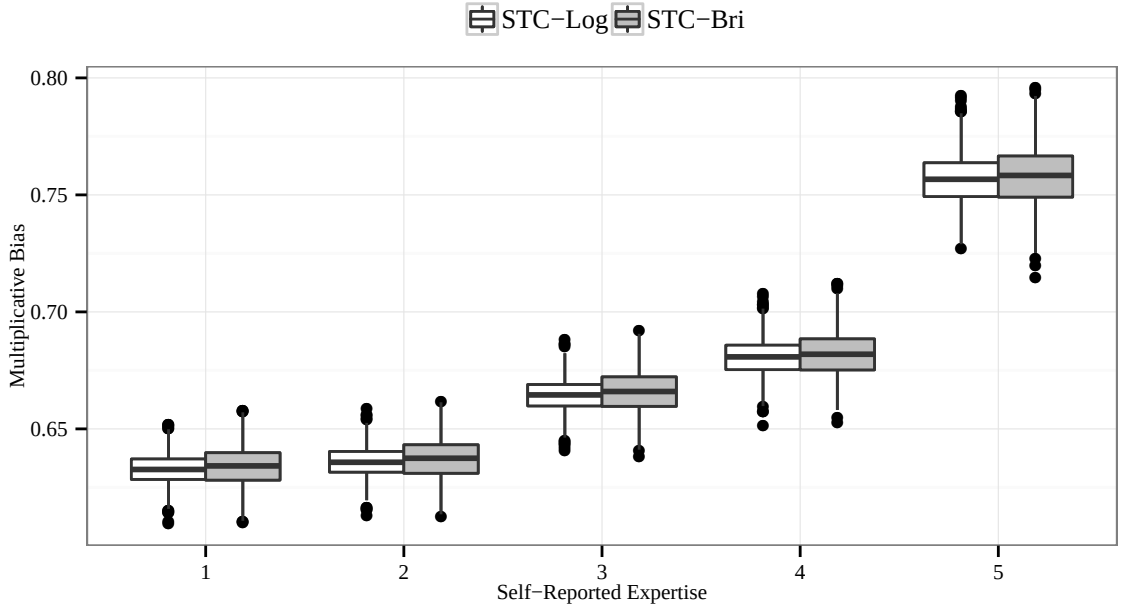


FIG 4. *Comparing the bias-levels across self-reported expertise under different approaches. Posterior distributions of* b_j *for* $j = 1, \ldots, 5$ *under the Sample-Then-Calibrate approach that optimizes over the logarithmic score and the Sample-Then-Calibrate approach that optimizes over the Brier score.*

7.4. *Group-Level Expertise Bias.* Recall from Section 2 that the experts were asked to self-assess their level of expertise (on a 1-to-5 scale with 1 = Not At All Expert to 5 = Extremely Expert) on any questions in which they participated. The self-reported expertise then divides the experts into 5 groups, with each group assigned a separate multiplicative bias term. This section uses the *Sample-Then-Calibrate*-procedure to explore the posterior distributions of these multiplicative bias terms. Figure 4 presents the posterior distributions of the bias terms with side-by-side box plots. Since the distributions fall completely below the *no-bias* reference-line at 1.0, all the groups are deemed under-confident.

The under-confidence, however, decreases as the level of expertise increases. For instance, the posterior probability that the most expert group is the least under-confident is approximately equal to 1.0, and the posterior probability of a strictly decreasing level of under-confidence is approximately 0.87. The latter probability is driven down by the inseparability of the two groups with the lowest levels of self-reported expertise. The fact that these groups are very similar suggests that the experts are poor at assessing how little they know about a subject that is strange to them. If these groups are combined into a single group, the posterior probability of a strictly decreasing level of under-confidence is approximately 1.0.

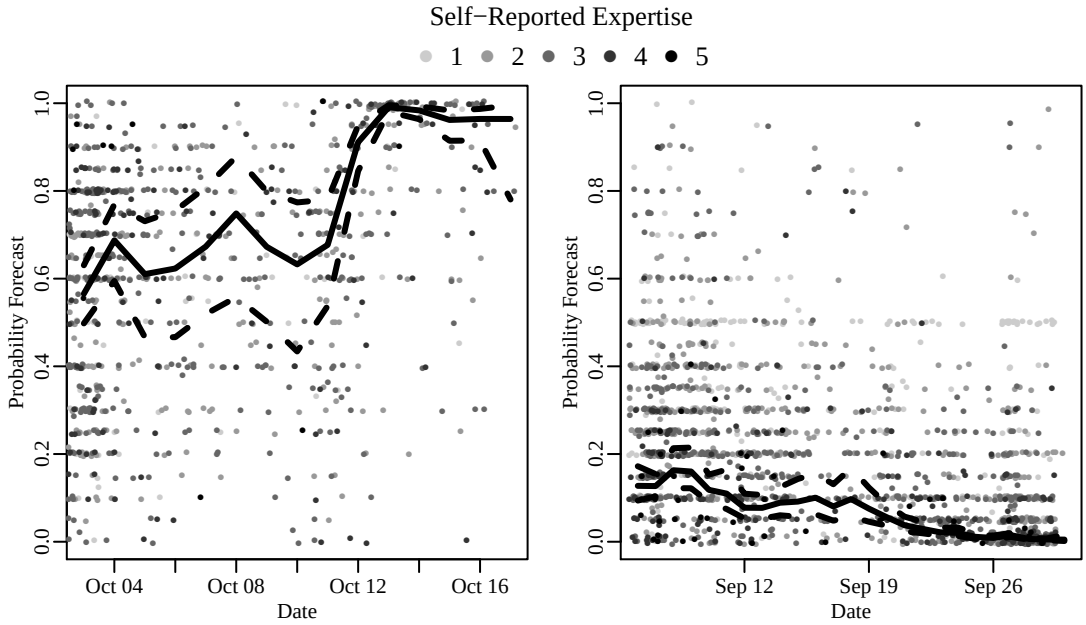
The decreasing trend in under-confidence can be reasoned by viewing the process of making a subjective probability as Bayesian updating: A completely ignorant expert aiming to minimize a reasonable loss function, such as the Brier score, has no reason to give anything but 0.5 as his probability forecast. However, as soon as the expert gains some knowledge about the event, he produces an updated forecast that is a compromise between his initial forecast and the new information acquired. The updated forecast is therefore conservative and too close to 0.5 as long as the expert remains only partially informed about the event. If most experts fall somewhere on this spectrum between ignorance and full information, their average forecast tends to fall strictly between 0.5 and the most-informed probability forecast (see Baron et al. (2013) for more details). Since expertise is to a large extent determined by subject-matter knowledge, the level of under-confidence can be expected to decrease as a function of the group's level of self-reported expertise.

Finding under-confidence in all the groups is a rather surprising result given that many previous studies have been conducted to show that experts are likely to be over-confident. For instance, Lichtenstein, Fischhoff and Phillips (1977); Morgan (1992); Bier (2004) summarize the results from numerous calibration studies, and conclude that experts are systematically over-condent about their probability assessments. Our result, however, is a statement about groups of experts and hence does not invalidate the possibility of the individual experts being overconfident. To make conclusions at the individual-level based on the group-level bias terms would be considered an *ecological*

inference fallacy (see, e.g., Lubinski and Humphreys (1996)).

7.5. *Question Difficulty and Other Measures.* One advantage of our model arises from its ability to produce estimates of interpretable question-specific parameters γ_k, σ_k^2, and τ_k^2. These quantities can be combined in many interesting ways to answer questions about different groups of experts or the questions themselves. For instance, being able to assess the difficulty of a question could lead to more principled ways of aggregating performance measures across questions or to novel insight on the kind of questions that are found difficult by experts (see, e.g., a discussion on the *Hard-Easy Effect* in Wilson and Wilson (1994)). To illustrate, recall that higher values of σ_k^2 suggest greater disagreement among the participating experts. Since experts are more likely to disagree over a difficult question than an easy one, it is reasonable to assume that σ_k^2 has a positive relationship with question difficulty. An alternative measure is given by τ_k that quantifies the volatility of the underlying circumstances that ultimately decide the outcome of the event. Therefore a high value of τ_k can cause the outcome of the event appear unstable and difficult to predict.

As a final illustration of our model, consider the two questions used for illustrative purposes in Section 2. Figure 5 is a copy of Figure 1 with the addition of a solid line surrounded by a dashed band. The solid line represents the posterior mean of the calibrated crowd belief as estimated by STC-Log. The dashed lines connect the point-wise 95% posterior intervals across different time points. Since $\hat{\sigma}_k^2 = 2.43$ and $\hat{\sigma}_k^2 = 1.77$ for the questions depicted in Figures 5(a) and 5(b), respectively, the first question provokes more disagreement among the experts than the second one. Intuitively this makes sense because the target event in Figure 5(a) is determined by several conditions that may change radically from one day to the next while the target event in Figure 5(b) is determined by a relatively steady stock market index. Therefore it is not surprising to find that the first question has $\hat{\tau}_k^2 = 1.38$ and the second one has $\hat{\tau}_k^2 = 0.198$. We conclude that the first question can be considered inherently more difficult than the second one.



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

(b) *Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?*

FIG 5. *Scatterplots of the probability forecasts given for two questions in our dataset. The shadings represents the self-reported expertise of the expert who provided the probability forecast. The solid line gives the posterior mean of the calibrated crowd belief as estimated by STC-Log. The surrounding dashed lines connect the point-wise 95% posterior intervals.*

8. Discussion. This paper began with an introduction of a rather unorthodox but nonetheless realistic time-series setting where probability forecasts are made very infrequently by a heterogeneous group of experts. The resulting data is too sparse to be modeled well with standard time-series methods. In response to this lack of appropriate modeling procedures, our work introduces an interpretable time-series model that incorporates self-reported expertise to capture a sharp and well-calibrated estimate of the crowd belief. The model estimation is performed in two steps: The first step, known as the *sampling step*, samples constrained versions of the model parameters via Gibbs sampling. The sampling is done from standard distributions with fast convergence to the target distribution (see Appendix A for technical details). The second step, known as the *calibra-*

tion step, uses a one-dimensional optimization procedure to transforms the constrained parameter values to their unconstrained counterparts. To the best of our knowledge, this procedure extends forecasting literature towards rather unexplored areas of probability aggregation.

8.1. *Summary of Findings.* The model was applied to an unusually large dataset on expert probability forecasts. The estimated crowd belief was found to be sharp and well-calibrated under both in- and out-of-sample settings. This has direct implications on predictive power and model interpretability. Firstly, the model was shown to outperform other probability aggregators in terms of forecasting ability. In particular, the model was deemed a well-balanced compromise that avoids over-fitting without being overly conservative. Secondly, the crowd belief was used as the no-bias reference point to study the bias among groups of experts with different levels of self-reported expertise. All the groups were found to be under-confident. The under-confidence, however, decreased as the level of self-reported expertise increased. This result is about groups of experts and hence does not conflict with the well-known result of the individuals being over-confident (see, e.g., Lichtenstein, Fischhoff and Phillips (1977); Morgan (1992); Bier (2004)). Besides making predictions or studying group-level bias, the model can be used to generate estimates of many problem-specific parameters. These quantities have clear interpretations and can be combined in many interesting ways to explore a range of hypotheses about different types of questions and expert behavior.

8.2. *Limitations and Directions for Future Research.* Our model pertains model parsimony while addressing the main challenges that arise from modeling sparse probability forecasting data. Therefore it can be viewed as a basis for many future extensions. To give some ideas, recall that most of the model parameters were assumed constant over time. It is intuitively reasonable, however, that these parameters behave differently during different time intervals of the question. For instance, the level of disagreement (represented by σ_k^2 in our model) among the experts can be expected to decrease towards the final time point when the question resolves. This hypothesis could

be explored by letting $\sigma^2_{t,k}$ evolve dynamically as a function of the previous term $\sigma^2_{t-1,k}$ and random noise.

If the experts were asked to provide personal information during the data collection process, it may be of interest to study the forecasting behavior of different kinds of experts. For instance, this paper modeled the bias separately within each expertise group. This, however, is by no means restricted to the study of bias or its relation to self-reported expertise. Different parameter dependencies could be constructed based on many other expert characteristic, such as gender, education, or specialty, to produce a range of novel insight on the forecasting behavior across different groups of experts. Furthermore, expert characteristics could be interacted with question types, such as economic, domestic, or international, to learn about the kind of experts who generally perform well on certain types of questions. This would be of interest to the decision-maker who could use the information as a basis for consulting only a high-performing subset of the available experts.

Other future directions could aim to remove some of the obvious limitations of our model. For instance, recall that the random components are assumed to follow a normal distribution. This is a strong assumption that may not always be justified. Logit-probabilities, however, have been modeled with the normal distribution before (see, e.g., Erev, Wallsten and Budescu (1994)). Furthermore, the normal distribution is a rather standard assumption in psychological models (see, e.g., signal-detection theory in Tanner Jr and Swets (1954)). A second limitation resides in the assumption that both the observed and hidden processes are expected to grow linearly. This assumption could be relaxed, for instance, by adding higher order terms to the model. A more complex model, however, is likely to sacrifice interpretability. Given that our model is able to detect very intricate patterns in the crowd belief (see Figure 5), compromising interpretability for the sake of facilitating non-linear growth is hardly necessary.

APPENDIX A: TECHNICAL DETAILS OF THE SAMPLING STEP

The Gibbs sampler (Geman and Geman (1984)) iteratively samples all the unknown parameters from their full-conditional posterior distributions one block of parameters at a time. Since this is performed under the constraint $b_3 = 1$ to ensure model identifiability, the constrained parameter estimates should be denoted with a trailing (1) to maintain consistency with earlier notation. For instance, the constrained estimate of γ_k should be denoted by $\hat{\gamma}_k(1)$ while the unconstrained estimate is denoted by $\hat{\gamma}_k$. For the sake of clarity, however, the constraint suffix is omitted in this section. Nonetheless, it is important to keep in mind that all the estimates in this section are constrained.

Sample $X_{t,k}$

The hidden states are sampled via the *Forward-Filtering-Backward-Sampling* (FFBS) algorithm that first predicts the hidden states using a Kalman Filter and then performs a backward sampling procedure that treats these predicted states as additional observations (see, e.g., Carter and Kohn (1994); Migon et al. (2005) for details on FFBS). More specifically, the first part, namely the Kalman Filter, is deterministic and consists of a predict and an update step. Given all the other parameters except the hidden states, the predict step for the kth question is

$$X_{t|t-1,k} = \gamma_k X_{t-1|t-1,k}$$
$$P_{t|t-1,k} = \gamma_k^2 P_{t-1|t-1,k} + \tau_k^2,$$

where the initial values, $X_{0|0,k}$ and $P_{0|0,k}$, are equal to 0 and 1, respectively. The update step is

$$e_{t,k} = Y_{i,t,k} - b_{i,k} X_{t|t-1,k}$$
$$S_{t,k} = \sigma_k^2 + b_{i,k}^2 P_{t|t-1,k}$$
$$K_{t,k} = P_{t|t-1,k} b_{i,k} S_{t,k}^{-1}$$
$$X_{t|t,k} = X_{t|t-1,k} + K_{t,k} e_{t,k}$$

$$P_{t|t,k} = (1 - K_{t,k}b_{i,k})P_{t|t-1,k},$$

where $b_{i,k}$ is the corresponding bias term for the ith expert in the kth question. The update step is repeated sequentially for each observation $Y_{i,t,k}$ given at time t. For each such repetition of the update step, the previous posterior values, $X_{t|t,k}$ and $P_{t|t,k}$, should be considered as the new prior values, $X_{t|t-1,k}$ and $P_{t|t-1,k}$. After running the Kalman Filter up to the final time point at $t = T_k$, the final hidden state is sampled from $X_{T_k,k} \sim \mathcal{N}(X_{T_k|T_k,k}, P_{T_k|T_k,k})$. The remaining states are obtained via the backward sampling that is performed in reverse from

$$X_{t-1,k} \sim \mathcal{N}\left(V\left(\frac{\gamma_k X_{t,k}}{\tau_k^2} + \frac{X_{t|t,k}}{P_{t|t,k}}\right), V\right),$$

where

$$V = \left(\frac{\gamma_k^2}{\tau_k^2} + \frac{1}{P_{t|t,k}}\right)^{-1}$$

This can be viewed as backward updating that considers the Kalman Filter estimates as additional observations at each given time point. If the observation $\boldsymbol{Y}_{t,k}$ is completely missing at time t, the update step is skipped and the state estimates are sampled from

$$\mathcal{N}\left(\gamma_k X_{t-1|t-1,k}, \gamma_k^2 P_{t-1|t-1,k} + \tau_k^2\right)$$

Sample $\boldsymbol{b}$ and σ_k^2

First, vectorize all the response vectors $\boldsymbol{Y}_{t,k}$ into a single vector denoted $\boldsymbol{Y}_k = \left[\boldsymbol{Y}_{1,k}^T, \ldots, \boldsymbol{Y}_{T_k,k}^T\right]^T$. Since each $\boldsymbol{Y}_{t,k}$ is matched with $X_{t,k}$ via the time index t, we can form a $|\boldsymbol{Y}_k| \times J$ design-matrix by letting $\boldsymbol{X}_k = \left[(\boldsymbol{M}_k X_{1,k})^T, \ldots, (\boldsymbol{M}_k X_{T_k,k})^T\right]^T$. Given that the goal is to borrow strength across questions by assuming a common bias vector $\boldsymbol{b}$, the parameter values must be estimated in parallel for each question such that the matrices $\boldsymbol{X}_k$ can be further concatenated into $\boldsymbol{X} = [\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T]^T$ during every iteration. Similarly, $\boldsymbol{Y}_k$ must be further vectorized into a vector $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_K^T]^T$. The question-specific variance terms are taken into account by letting

$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2 \mathbf{1}_{1\times T_1}, \ldots, \sigma_K^2 \mathbf{1}_{1\times T_K})$. After adopting the non-informative prior $p(\boldsymbol{b}, \sigma_k^2 | \boldsymbol{X}_k) \propto \sigma_k^{-2}$ for each $k = 1, \ldots, K$, the bias vector are sampled from

$$\boldsymbol{b}|\ldots \;\sim\; \mathcal{N}_J\left((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}, (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\right) \tag{6}$$

Since the covariance matrix in Equation (6) is diagonal, the constraint is enforced at this point by letting $b_3 = 1$. The variance parameter is then sampled from

$$\sigma_k^2|\ldots \;\sim\; \text{Inv-}\chi^2\left(|\boldsymbol{Y}_k| - J, \frac{1}{|\boldsymbol{Y}_k| - J}(\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{b})^T(\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{b})\right),$$

where the distribution is a scaled inverse-χ^2 (see, e.g., Gelman et al. (2003)). Since the experts are not required to give a new prediction at every time unit, the design matrices must be trimmed accordingly such that their dimensions match up with the dimensions of the observed matrices.

Sample γ_k and τ_k^2

Estimating the parameters related to the hidden process are estimated via a regression setup. More specifically, after adopting the non-informative prior $p(\gamma_k, \tau_k^2|\boldsymbol{X}_k) \propto \tau_k^{-2}$, the parameter values are sampled from

$$\begin{aligned}
\gamma_k|\ldots \;&\sim\; \mathcal{N}\left(\frac{\sum_{t=2}^{T_k} X_{t,k}X_{t-1,k}}{\sum_{t=1}^{T_k-1} X_{t,k}^2}, \frac{\tau_k^2}{\sum_{t=1}^{T_k-1} X_{t,k}^2}\right) \\
\tau_k^2|\ldots \;&\sim\; \text{Inv-}\chi^2\left(T_k - 1, \frac{1}{T_k - 1}\sum_{t=2}^{T_k}(X_{t,k} - \gamma_k X_{t-1,k})^2\right),
\end{aligned}$$

where the final distribution is a scaled inverse-χ^2 (see, e.g., Gelman et al. (2003)).

REFERENCES

ALLARD, D., COMUNIAN, A. and 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., WALLSTEN, T. S. and ZAUBERMAN, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied* **6** 130-147.

BAARS, J. A. and MASS, C. F. (2005). Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics. *Weather and forecasting* **20** 1034–1047.

BARON, J., UNGAR, L. H., MELLERS, B. A. and E., T. P. (2013). Two reasons to make aggregated probability forecasts more extreme. *submitted.*

BIER, V. (2004). Implications of the research on expert overconfidence and dependence. *Reliability Engineering & System Safety* **85** 321–329.

BONFERRONI, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilitá. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8** 3-62.

BRIER, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* **78** 1-3.

BRÖCKER, J. and SMITH, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting* **22** 651–661.

BUJA, A., STUETZLE, W. and SHEN, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*.

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553.

CHEN, Y. (2009). Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets. *Department of Computer Science Iowa State University*.

CLEMEN, R. T. and WINKLER, R. L. (2007). Aggregating probability distributions. *Advances in Decision Analysis* 154–176.

COOKE, R. M. (1991). Experts in uncertainty: opinion and subjective probability in science.

EREV, I., WALLSTEN, T. S. and BUDESCU, D. V. (1994). Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes. *Psychological Review* **66** 519-527.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003). *Bayesian data analysis*. CRC press.

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 1360–1383.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **6** 721–741.

GENEST, C. and ZIDEK, J. V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science* **1** 114-148.

GENT, I. P. and WALSH, T. (1996). Phase transitions and annealed theories: Number partitioning as a case study'. In *ECAI* 170–174. Citeseer.

GNEITING, T., STANBERRY, L. I., GRIMIT, E. P., HELD, L. and JOHNSON, N. A. (2008). Rejoinder on: Assessing

probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17** 256–264.

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

HASTINGS, C., MOSTELLER, F., TUKEY, J. W. and WINSOR, C. P. (1947). Low moments for small samples: a comparative study of order statistics. *The Annals of Mathematical Statistics* **18** 413–426.

HAYES, B. (2002). The easiest hard problem. *American Scientist* **90** 113–117.

KARMARKAR, N. and KARP, R. M. (1982). *The differencing method of set partitioning*. Computer Science Division (EECS), University of California Berkeley.

KELLERER, H., PFERSCHY, U. and PISINGER, D. (2004). *Knapsack problems*. Springer.

LICHTENSTEIN, S., FISCHHOFF, B. and PHILLIPS, L. D. (1977). *Calibration of probabilities: The state of the art*. Springer.

LUBINSKI, D. and HUMPHREYS, L. G. (1996). Seeing the forest from the trees: When predicting the behavior or status of groups, correlate means. *Psychology, Public Policy, and Law* **2** 363.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics* **21** 1087.

MIGON, H. S., GAMERMAN, D., LOPES, H. F. and FERREIRA, M. A. (2005). Dynamic models. *Handbook of Statistics* **25** 553–588.

MILLS, T. C. (1991). *Time series techniques for economists*. Cambridge University Press.

MORGAN, M. G. (1992). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press.

MURPHY, A. H. and WINKLER, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review* **115** 1330–1338.

NEAL, R. M. (2003). Slice sampling. *Annals of statistics* 705–741.

NICULESCU-MIZIL, A. and CARUANA, R. (2005). Obtaining Calibrated Probabilities from Boosting. In *UAI* 413.

PEPE, M. S. (2003). The statistical evaluation of medical tests for classification and prediction.

PLATT, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **10** 61–74.

PRIMO, C., FERRO, C. A., JOLLIFFE, I. T. and STEPHENSON, D. B. (2009). Calibration of probabilistic forecasts of binary events. *Monthly Weather Review* **137** 1142–1149.

RAFTERY, A. E., GNEITING, T., BALABDAOUI, F. and POLAKOWSKI, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133** 1155–1174.

RANJAN, R. and GNEITING, T. (2010). Combining Probability Forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 71-91.

SANDERS, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology* **2** 191–201.

SATOPÄÄ, V. A., BARON, J., FOSTER, D. P., MELLERS, B. A., TETLOCK, P. E. and UNGAR, L. H. (2013). Combining Multiple Probability Predictions Using a Simple Logit Model. *submitted*.

SHLYAKHTER, A. I., KAMMEN, D. M., BROIDO, C. L. and WILSON, R. (1994). Quantifying the credibility of energy projections from trends in past data: The US energy sector. *Energy Policy* **22** 119–130.

TANNER JR, W. P. and SWETS, J. A. (1954). A decision-making theory of visual detection. *Psychological review* **61** 401.

TETLOCK, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

UNGAR, L., MELLERS, B., SATOPÄÄ, V., TETLOCK, P. and BARON, J. (2012). The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions. In *2012 AAAI Fall Symposium Series*.

VISLOCKY, R. L. and FRITSCH, J. M. (1995). Improved model output statistics forecasts through model consensus. *Bulletin of the American Meteorological Society* **76** 1157–1164.

WALLACE, B. C. and DAHABREH, I. J. (2012). Class probability estimates are unreliable for imbalanced data (and how to fix them). In *Data Mining (ICDM), 2012 IEEE 12th International Conference on* 695–704. IEEE.

WALLSTEN, T. S., BUDESCU, D. V. and EREV, I. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making* **10** 243–268.

WILSON, A. G. and WILSON, A. G. (1994). Cognitive Factors Affecting Subjective Probability Assessment.

WILSON, P. W., DAGOSTINO, R. B., LEVY, D., BELANGER, A. M., SILBERSHATZ, H. and KANNEL, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* **97** 1837–1847.

WINKLER, R. L. and JOSE, V. R. R. (2008). Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17** 251–255.

WINKLER, R. L. and MURPHY, A. H. (1968). Good Probability Assessors1.

WRIGHT, G., ROWE, G., BOLGER, F. and GAMMACK, J. (1994). Coherence, calibration, and expertise in judgmental probability forecasting. *Organizational Behavior and Human Decision Processes* **57** 1–25.

PHILADELPHIA, PA 19104- 6340, USA

E-MAIL: satopaa@wharton.upenn.edu