Judgment and Decision Making (2023), **Vol. 18**:e40 1–20
doi:10.1017/jdm.2023.39

THEORY ARTICLE

A normative model for Bayesian combination of subjective probability estimates

Susanne Trick[1,2], Constantin A. Rothkopf[1,2,3], and Frank Jäkel[1,2]

[1]Centre for Cognitive Science, Technical University of Darmstadt, Darmstadt, Germany; [2]Institute of Psychology, Technical University of Darmstadt, Darmstadt, Germany and [3]Frankfurt Institute for Advanced Studies, Goethe University, Frankfurt, Germany

Corresponding author: Susanne Trick; Email: susanne.trick@tu-darmstadt.de

Received: 18 April 2023; **Revised:** 11 September 2023; **Accepted:** 11 September 2023

Keywords: forecast aggregation; normative model; Bayesian inference; calibration; confidence

Abstract
Combining experts' subjective probability estimates is a fundamental task with broad applicability in domains ranging from finance to public health. However, it is still an open question how to combine such estimates optimally. Since the beta distribution is a common choice for modeling uncertainty about probabilities, here we propose a family of normative Bayesian models for aggregating probability estimates based on beta distributions. We systematically derive and compare different variants, including hierarchical and non-hierarchical as well as asymmetric and symmetric beta fusion models. Using these models, we show how the beta calibration function naturally arises in this normative framework and how it is related to the widely used Linear-in-Log-Odds calibration function. For evaluation, we provide the new Knowledge Test Confidence data set consisting of subjective probability estimates of 85 forecasters on 180 queries. On this and another data set, we show that the hierarchical symmetric beta fusion model performs best of all beta fusion models and outperforms related Bayesian fusion models in terms of mean absolute error.

1. Introduction

Experts' subjective probability estimates of an event's occurrence or the correctness of a statement are of particular importance in many different domains such as finance, business, marketing, politics, engineering, meteorological, ecological, and environmental science, and public health (McAndrew et al., 2021). While statistical models are usually limited in applicability by requiring sufficiently large and complete data sets, human forecasts can overcome this limitation taking advantage of human experience and intuition (Clemen, 1989; Clemen & Winkler, 1986; Genest & Zidek, 1986). The probability estimates can be given either as forecasts of events, e.g., rain probabilities in meteorological science or probabilities for the outcomes of geopolitical events such as elections (Graefe, 2018; Turner et al., 2014), other binary classifications, or the quantification of the experts' confidence on a prediction or the answer to a specific question (Karvetski et al., 2013; Prelec et al., 2017).

We assume humans to have internal beliefs, which are expressed as the subjective probability estimates they provide. However, we do not necessarily assume that they compute their beliefs by

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

doing Bayesian inference in their heads (Griffiths & Tenenbaum, 2006; Lee, 2018a). Subjective probability estimates provided by humans are often miscalibrated meaning that they are overconfident or underconfident. A well-calibrated forecaster's probability estimates match the respective relative frequency of occurrence, i.e., $100x\%$ of the answers for which the forecaster predicts probability x are correct. In contrast, a miscalibrated forecaster's probability estimate is more (overconfidence) or less (underconfidence) extreme (Morgan, 2014). Measuring the calibration of forecasters, quantifying it with a calibration function, and recalibrating their given probabilities using this calibration function can improve forecasts (Graham, 1996).

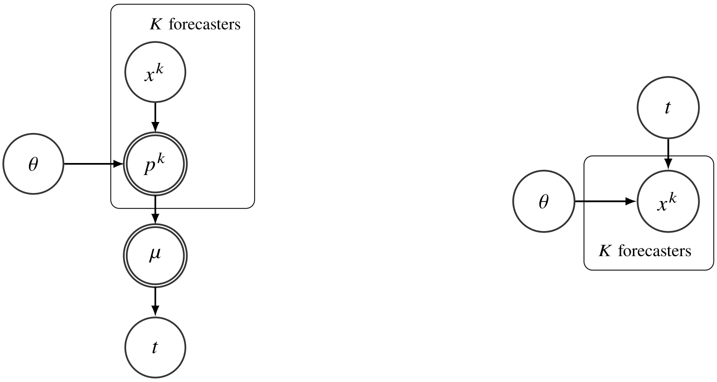
Compared to individual forecasts, combining forecasts usually increases performance (Budescu & Chen, 2015; McAndrew et al., 2021; Satopää, 2022; Turner et al., 2014), since a group of forecasters provides more information than a single forecaster (Clemen & Winkler, 1999). A distinction is made between behavioral and mathematical aggregation of forecasts. While behavioral aggregation can be subsumed as a process in which the forecasters negotiate a consensus (Minson et al., 2018; Silver et al., 2021), mathematical aggregation involves mathematical rules or models to combine individual forecasts to an aggregated forecast (Clemen & Winkler, 1999; Hanea et al., 2021; Wilson, 2017). In this work, we focus on mathematical aggregation.

A popular choice for mathematical aggregation of probability estimates are linear opinion pools. While unweighted linear opinion pools, i.e., simple averages, often perform surprisingly well (Turner et al., 2014), numerous weighted linear opinion pools have been designed, because not all opinions are necessarily of the same value. The weights can be selected based on the forecasters' performance (Budescu & Chen, 2015; Cooke, 1991; Hanea et al., 2021), the coherence of their answers (Karvetski et al., 2013), or by the number of cues available to them (Budescu & Rantilla, 2000), or can be optimized for maximum performance (Ranjan & Gneiting, 2010). In the latter approach by Ranjan and Gneiting (2010), they additionally calibrate the weighted linear opinion pool by transforming it with the cumulative distribution function of the beta distribution. To deal with under- and overconfidence of linear opinion pools, also trimmed opinion pools (Grushka-Cockayne et al., 2017) or methods for (anti-) extremizing linear opinion pools (Baron et al., 2014; Lichtendahl Jr. et al., 2022) have been introduced. In addition to linear pooling, there are also multiplicative pooling methods, e.g., Independent Opinion Pool (Berger, 1985), which is a renormalized product of the forecasts, or logarithmic or geometric pooling (Berger, 1985; Dietrich & List, 2016), which is a weighted product of forecasts. Both, linear and multiplicative pooling methods can also be used with transformations of the forecasts, e.g., as a probit average (Satopää et al., 2023) or a geometric mean of odds (Satopää et al., 2014).

Although, as seen above, there are already many different methods for combining probability forecasts, according to the review by McAndrew et al. (2021), an open challenge is a 'normative theory for how to combine expert opinions into a single consensus distribution' (McAndrew et al., 2021). Therefore, in this paper, we propose a normative model for combining probability forecasts that models the behavior of individual forecasters and fuses them accordingly.

Several Bayesian statistical models for combining subjective probability estimates have recently been introduced (Hanea et al., 2021; Lee & Danileiko, 2014; Satopää, 2022; Turner et al., 2014). Lee and Danileiko (2014) ask the forecasters for percentages or probabilities (e.g., 'What percentage of the world's water is not freshwater?') and thus consider probabilities as ground truth. Their model is unsupervised, so they do not use any historical seed queries from which the forecasters' behavior can be learned. Hanea et al. (2021) and Satopää (2022) consider binary truth values as ground truth, but also work with unsupervised models. In contrast, Turner et al. (2014), who also assume binary ground truth values, propose supervised models for combining probability forecasts. In particular, they investigate whether it is better first to calibrate the given probabilities and then average the recalibrated probability estimates or first to average them and recalibrate the resulting average. For calibration, they use the Linear-in-Log-Odds (LLO) calibration function. Turner et al. (2014) compare non-hierarchical models and hierarchical models and models based on probability or log-odds. They evaluated all of these models on one data set consisting of a total of 11,551 answers to 176 geopolitical questions, and conclude that first recalibrating and then combining the recalibrated probability estimates results

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press



(a) Calibrate then Average Model (Turner et al., 2014)

(b) Normative Fusion Model

Figure 1. *Graphical models of one of the models proposed by Turner et al. (2014), Calibrate then Average, (a) and an exemplary normative fusion model (b). The models are simplified to the forecasts x^k of K forecasters to only one query with truth value t.*

in the best performance. Although their approach uses Bayesian inference to infer the best parameters for different combination rules, the combination rules themselves are not motivated normatively. In Figure 1(a), we show a simplified graphical model of one of the fusion models presented by Turner et al. (2014), Calibrate then Average. In this model, the forecasts x^k of K forecasters are calibrated using a deterministic calibration function with parameters θ (which is the LLO function in their model) to obtain the calibrated forecasts p^k. These calibrated forecasts p^k are then fused to the fused forecast μ using a deterministic fusion method, i.e., averaging. The truth value t is drawn from a Bernoulli distribution with parameter μ. Note that Turner's approach thus models how the truth value t is generated from the forecast data x^k, as it is usually done in a discriminative model.

In contrast, a normative fusion model as we propose it in this work (Figure 1(b)) expresses how the true value t generates the forecasts x^k, i.e., the data-generating process, in a generative model. The forecasts x^k are generated from some probability distribution conditioned on the true label t with the respective distribution parameters θ. Thus, after learning the parameters θ from labeled training data, the model represents the forecasting behavior of the forecasters conditioned on t. In particular, as was shown for a model combining classifier outputs (Trick & Rothkopf, 2022), it models the forecasters' bias, variance, and uncertainty. In addition, the normative fusion model implicitly calibrates the forecasts without the need of an explicit calibration function. New forecasts x^k are fused using Bayes' rule by inferring the posterior probability of t given the forecasts x^k and the learned model parameters θ. This is normative fusion behavior, because Bayesian inference is normative. Of course, the parameters θ can also be modeled for each forecaster individually as θ^k, which allows modeling each forecaster's individual forecasting behavior conditioned on t.

Lindley (1985) proposed such a normative model for combining probability estimates (nicely explained by Jacobs, 1995). He transforms the probabilities to log-odds and models these log-odds with Gaussian distributions conditioned on the truth value t that indicates whether the respective event occurred or not. For fusing the predictions, the posterior probability of t given the learned Gaussian models and the predictions to be fused needs to be inferred using Bayes' rule.

Another natural way to model the combination of probability estimates normatively is to model the probabilities directly with a beta distribution without the need of any transformation. As far as we know, a Bayesian model for combining human probability estimates using beta distributions has not been worked out in detail yet, but it has been mentioned in an example in the book of Berger (1985). Steyvers et al. (2014) modeled probability forecasts with beta distributions, however, not for fusion or calibration of probability estimates but for evaluating forecast performances using ROC curves. Here, we propose a normative model for combining probability estimates that models the probabilities with a beta distribution conditioned on the true label t of each forecast.

In this vein, we will show that modeling probabilistic forecasts with beta distributions conditioned on the true label t implicitly calibrates them. This calibration function named beta calibration has recently been introduced in a machine learning context (Ji et al., 2020; Kull et al., 2017). The LLO calibration function, used by Turner et al. (2014) and Lee and Danileiko (2014), can be shown to be a special case of the beta calibration function.

Turner et al. (2014) evaluate their Bayesian fusion models on the The Good Judgment data set by the IARPA Aggregative Contingency Estimation (ACE) System, which is the most popular data set for evaluating forecast aggregation methods and is used for evaluation in many approaches on forecast aggregation (Budescu & Chen, 2015; Hanea et al., 2021; Satopää, 2022; Steyvers et al., 2014; Turner et al., 2014; Wang et al., 2021). It includes questions on the probability of future geopolitical events, such as the outcome of elections. Turner et al. (2014) only evaluated on a subset of this data set including only binary events and only similarly framed questions.

This data set used by Turner et al. (2014) consists of 176 events or questions, which are answered by 1,290 forecasters. While the high number of events seems to be beneficial for modeling the forecasters' behavior, all forecasters only provided forecasts for a subset of these 176 questions. On average, a forecaster provides only 8.25 answers and 221 (169/122) forecasters only provide 1 (2/3) answer. These low numbers of forecasts per person is unfavorable for modeling the behavior of individual forecasters. Also, modeling their behavior conditioned on whether the event has occurred or not is difficult with this data set since only 37 out of 176 events are positive events that actually occurred.

Several other data sets consisting of probability estimates of multiple forecasters also show similar drawbacks. The ACE-IDEA data set (Hanea et al., 2021) includes forecasts on 155 events, but on average each forecaster only replied to about 19 queries. Other data sets consist of less queries to be predicted or answered (Graefe, 2018; Hanea et al., 2021; Karvetski et al., 2013; Prelec et al., 2017), of which the highest number of queries is about 80 (Prelec et al., 2017). However, 80 answers per forecaster is still a small number for modeling the forecasters' behavior, particularly if we divide the data into training and test sets and model the answers to true/false queries separately. Also, this data set is not publicly available.

Since the available data sets do not provide much information about single forecasters, in this work, we publish a new data set consisting of 180 true and false statements, for each of which 85 forecasters provide their confidence on the statement's correctness.

Thus, the contribution of this paper is four-fold: First, we present a family of natural normative models for aggregating probability forecasts based on the beta distribution. Second, we introduce beta calibration in expert fusion contexts and show the connection between the widely used LLO calibration function and the beta distribution. Third, for evaluating our normative models, we provide a new data set consisting of subjective probability forecasts that includes a sufficient number of data points to model the behavior of individual forecasters. Fourth, we systematically evaluate the proposed normative beta fusion models on a data set by Turner et al. (2014) and our new data set, compare them to the models proposed in the work of Turner et al. (2014), and provide some general findings about the respective performance of the considered models regarding several different scores.

2. Modeling probability estimates with beta distributions

In the following, we assume that K human subjects (forecasters) provide probability estimates (forecasts) on N binary queries. These queries can be questions about future events, e.g., 'Will XY happen?', factual statements, or other binary classifications, and must have in common that they can be answered with 'true' or 'false'. The forecasters provide a probability for each item, which can be interpreted as either their belief that the respective event will occur or their confidence in their answer's

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

correctness. In the latter case, a probability of 0 means that the forecaster is 100% certain that the correct answer is ‘false’, while a probability of 1 indicates that the forecaster is 100% certain the answer is ‘true’.

The forecast given by forecaster k for query n is formalized as x_n^k with $n = 1, \dots, N$, $k = 1, \dots, K$. The true label t_n for query n, which is the ground truth, can take values 0 or 1, where 0 indicates truth value ‘false’ and 1 indicates truth value ‘true’. We assume the forecasts x_n^k to be conditionally independent given the true label t_n. For a discussion of this assumption, which does not hold in general, we refer to Section 4.

The beta distribution is the natural choice for modeling proportions and probabilities, because it is the standard distribution for probabilities in Bayesian statistics and the conjugate prior of the Bernoulli distribution. Therefore, here we model the forecasts x_n^k with a beta distribution conditioned on the true label $t_n = j, j \in \{0, 1\}$. We thus assume that humans have some skill to differentiate between true and false queries.

Whereas usually the beta distribution is parameterized with two shape parameters α and β, in the following, we will parameterize it alternatively with mean μ and the proportion ρ of the maximum possible variance given μ, which is $\mu(1-\mu)$. This reparametrization increases the parameters’ interpretability and computationally improves the performance of Gibbs sampling by reducing correlations between the variables. Thus, the forecasts x_n^k are modeled with a beta distribution conditioned on the true label $t_n = j, j \in \{0, 1\}$, with parameter μ_j^k as the beta distribution’s mean and parameter ρ_j^k as the proportion of the beta distribution’s maximum variance,

$$x_n^k | t_n = j \sim \text{Beta}'(\mu_j^k, \rho_j^k), \tag{2.1}$$

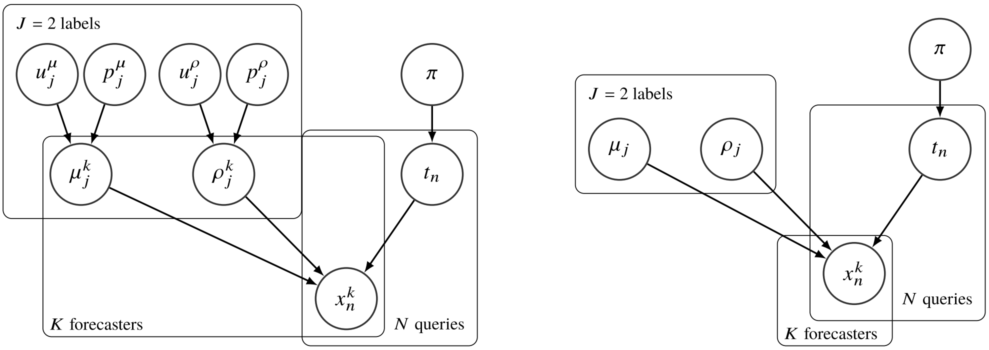
where $\text{Beta}'(\mu_j^k, \rho_j^k)$ is identical to $\text{Beta}(\alpha_j^k, \beta_j^k)$ with

$$\begin{aligned} \alpha_j^k &= \mu_j^k \eta_j^k, \\ \beta_j^k &= (1 - \mu_j^k)\eta_j^k, \\ \eta_j^k &= \frac{\mu_j^k(1 - \mu_j^k)}{\nu_j^k} - 1, \\ \nu_j^k &= \rho_j^k \mu_j^k (1 - \mu_j^k), \end{aligned} \tag{2.2}$$

and ν_j^k is the beta distribution’s variance. The true label t_n is modeled with a Bernoulli distribution with parameter π.

2.1. Hierarchical Beta Fusion Model

In many cases, forecasters provide their forecasts to only a subset of the available queries. In order to also be able to accurately model the forecasting behavior of forecasters who only provided a small number of forecasts, we propose the Hierarchical Beta Fusion Model. This allows taking advantage of the statistical properties of hierarchical models and their psychological interpretations. Their statistical properties are increased statistical power since parameter inference is based on more data that share information across groups, and reduced variance of parameter estimates, known as shrinkage (Britten et al., 2021). In addition, they allow modeling of individual differences between forecasters, make it possible to model forecasters who only provided few forecasts, and provide information on the distribution of the forecasters’ individual parameters (Lee, 2018b). In particular, the model’s hyperparameters indicate how the forecasters behave on average, how variable their behavior is, and the associated hyperpriors make assumptions about the forecasters’ average behavior and variability explicit.



(a) Hierarchical Beta Fusion Model

(b) Non-Hierarchical Beta Fusion Model

Figure 2. *Graphical models of the hierarchical (a) and non-hierarchical (b) beta fusion models.*

In the Hierarchical Beta Fusion Model, we use a beta prior on the model parameter μ_j^k, parameterized with mean u_j^μ and maximum variance proportion p_j^μ. The proportion of the maximum variance ρ_j^k is also modeled with a beta distribution with mean u_j^ρ and maximum variance proportion p_j^ρ. To avoid values of ρ_j^k too close to 0 or 1 that may get the Gibbs sampler in trouble, we constrained this beta distribution and all other beta priors on maximum variance proportions ρ, p between 0.001 and 0.999. As prior distribution for the proportion π of true queries, we chose a uniform beta distribution Beta(1,1). The graphical model of the Hierarchical Beta Fusion Model is shown in Figure 2(a). A complete overview over the corresponding modeling distributions and priors, also including the priors of the hyperparameters $u_j^\mu, p_j^\mu, u_j^\rho, p_j^\rho$, is given as

$$\begin{aligned} t_n &\sim \text{Bernoulli}(\pi) \\ x_n^k | t_n = j &\sim \text{Beta}'(\mu_j^k, \rho_j^k) \\ \mu_j^k &\sim \text{Beta}'(u_j^\mu, p_j^\mu) \\ \rho_j^k &\sim \text{Beta}'(u_j^\rho, p_j^\rho) \end{aligned} \qquad \begin{aligned} \pi &\sim \text{Beta}(1,1) \\ u_j^\mu &\sim \text{Beta}(1,1) \\ p_j^\mu &\sim \text{Beta}(1,1) \\ u_j^\rho &\sim \text{Beta}(1,1) \\ p_j^\rho &\sim \text{Beta}(1,1). \end{aligned} \tag{2.3}$$

Based on labeled training data, the model parameters μ_j^k, ρ_j^k, and π can be inferred using Gibbs sampling (e.g., by specifying this model in JAGS). Since we assume that human forecasters are on average consistent between different queries, we can use the learned parameters to infer the posterior probability of the true label t_n of new unseen forecasts x_n^k as the fusion result. Inference of t_n can either also be realized using Gibbs sampling, or the posterior probability of t_n can be computed analytically using the closed-form probability density function of the beta distribution:

$$p(t_n = j | \boldsymbol{x_n}, \boldsymbol{\mu_j}, \boldsymbol{\rho_j}, \pi) \propto \pi^j (1-\pi)^{1-j} \prod_{k=1}^{K} \text{Beta}'(x_n^k; \mu_j^k, \rho_j^k). \tag{2.4}$$

2.2. Non-hierarchical Beta Fusion Model

The normative Hierarchical Beta Fusion model represents the forecasting behavior of each forecaster individually, i.e., for each forecaster an individual set of beta parameters μ_j^k, ρ_j^k is learned. In this way, we can model interindividual differences between forecasters, e.g., different levels of expertise, and can exploit these learned properties for fusion by giving more weight to a better-performing forecaster. However, since the related approach by Turner et al. (2014) also compared hierarchical and non-hierarchical versions of their fusion models, we also compare our normative Hierarchical Beta Fusion Model to a non-hierarchical version of the model, which assumes exchangeable forecasters that behave similarly. In this non-hierarchical Beta Fusion Model, we model the forecasts x_n^k with the same beta distributions for all forecasters with shared parameters μ_j and ρ_j for $k = 1, \dots, K$ and $j \in \{0, 1\}$. Thus, we learn only two beta distributions for all forecasters, the beta distribution with mean μ_0 and maximum variance proportion ρ_0 models the forecasting behavior of all forecasters for false queries, the beta distribution with mean μ_1 and maximum variance proportion ρ_1 models their forecasting behavior for true queries. The priors on μ_j and ρ_j are uniform distributions Beta(1,1). As for the Hierarchical Beta Fusion Model in Section 2.1, the prior for proportion π is an uninformative beta prior Beta(1,1). We illustrate the graphical model of the non-hierarchical Beta Fusion Model in Figure 2(b). All corresponding modeling distributions and priors can be summarized as

$$\begin{aligned} t_n &\sim \text{Bernoulli}(\pi) \\ x_n^k | t_n = j &\sim \text{Beta}'(\mu_j, \rho_j) \end{aligned} \qquad \begin{aligned} \mu_j &\sim \text{Beta}(1,1) \\ \rho_j &\sim \text{Beta}(1,1) \\ \pi &\sim \text{Beta}(1,1). \end{aligned} \tag{2.5}$$

Again, given labeled training data, we can infer the model parameters μ_j, ρ_j, and π using Gibbs sampling, and the fused result for some unseen forecasts x_n^k of multiple forecasters is the posterior probability over t_n given the learned model parameters and forecasts x_n^k. Using equation (2.4) as for the Hierarchical Beta Fusion Model and dropping index k for μ and ρ allows the analytical computation of the posterior over t_n.

2.3. Beta calibration

A forecaster is well calibrated if their probability estimate matches the respective relative frequency of occurrence, i.e., if $100x\%$ of the statements to which the forecaster assigns a probability of x are true or $100x\%$ of the events to which the forecaster assigns a probability of x occur (Brenner et al., 1996). The calibration of a human forecaster can be measured empirically by binning the provided probability estimates and computing the proportions of true events for each bin. It is customary to illustrate this relationship with the so-called calibration curve, which plots the proportions of true events as a function of the human forecast probabilities. If the resulting calibration curve is the identity function, the forecaster is perfectly calibrated. If not, a function can be fitted to the empirical calibration curve. This calibration function can then also be used to recalibrate probability estimates, i.e., to correct for overconfident or underconfident judgments (Turner et al., 2014).

While various different functions can serve as calibration functions, the LLO function is frequently used. For example, Lee and Danileiko (2014) and Turner et al. (2014) explicitly include the LLO function in their Bayesian fusion models for calibrating the provided forecasts. However, by modeling the provided forecasts with a probability distribution, one can also calibrate them implicitly. This means that given a probabilistic generative fusion model, as we provide in this work, the calibration function is not chosen empirically, but instead the normative calibration function for the respective model can be derived.

With the Hierarchical Beta Fusion Model, we model forecasts x_n^k of forecaster k conditioned on the true label t_n with a beta distribution

$$P(x_n^k = x | t_n = j) = \frac{x^{\alpha_j^k - 1}(1-x)^{\beta_j^k - 1}}{\mathrm{B}(\alpha_j^k, \beta_j^k)}. \tag{2.6}$$

The corresponding calibration function, which is called the beta calibration function, can be derived using Bayes' rule

$$\begin{aligned}\mathrm{BC}(x) = P(t_n = 1 | x_n^k = x) &= \frac{\frac{x^{\alpha_1^k-1}(1-x)^{\beta_1^k-1}}{\mathrm{B}(\alpha_1^k,\beta_1^k)}\pi}{\frac{x^{\alpha_1^k-1}(1-x)^{\beta_1^k-1}}{\mathrm{B}(\alpha_1^k,\beta_1^k)}\pi + \frac{x^{\alpha_0^k-1}(1-x)^{\beta_0^k-1}}{\mathrm{B}(\alpha_0^k,\beta_0^k)}(1-\pi)} \\ &= \frac{1}{1 + \frac{\mathrm{B}(\alpha_1^k,\beta_1^k)}{\mathrm{B}(\alpha_0^k,\beta_0^k)} \frac{(1-x)^{\beta_0^k-\beta_1^k}}{x^{\alpha_1^k-\alpha_0^k}} \frac{1-\pi}{\pi}}\end{aligned} \tag{2.7}$$

with $\pi = P(t_n = 1)$ as introduced above. The function in (2.7) has been first introduced by Kull et al. (2017) in the context of machine learning for calibrating the probabilistic outputs of classification algorithms.

Interestingly, the LLO calibration function used by Turner et al. (2014) and Lee and Danileiko (2014) can be derived as a special case of the beta calibration function (Kull et al., 2017). If we constrain the beta distributions to be symmetric around $\frac{1}{2}$, i.e., $\alpha = \alpha_0^k = \beta_1^k$ and $\beta = \beta_0^k = \alpha_1^k$, the resulting calibration function is

$$\begin{aligned}\mathrm{LLO}(x) = P(t_n = 1 | x_n^k = x) &= \frac{\frac{\pi}{1-\pi} x^{\beta-\alpha}}{\frac{\pi}{1-\pi} x^{\beta-\alpha} + (1-x)^{\beta-\alpha}} \\ &= \frac{\delta x^{\gamma}}{\delta x^{\gamma} + (1-x)^{\gamma}}\end{aligned} \tag{2.8}$$

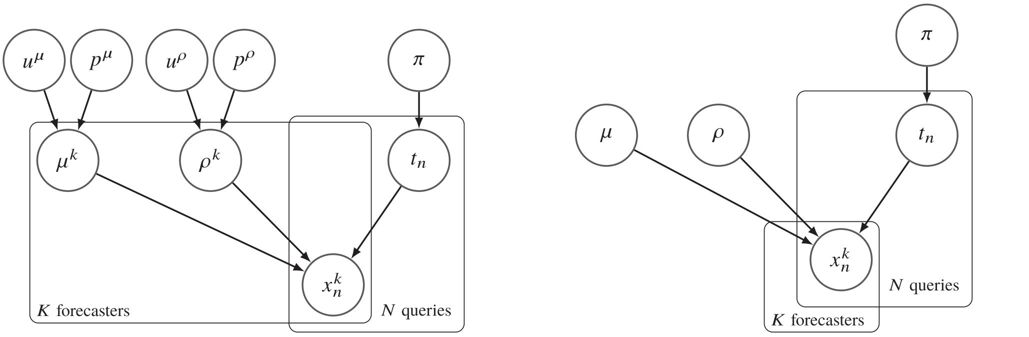
with $\delta = \frac{\pi}{1-\pi}$ and $\gamma = \beta - \alpha$. Beta calibration in (2.7) is more flexible than LLO calibration in (2.8) because it does not assume symmetric beta distributions for $t_n = 0$ and $t_n = 1$. Thus, beta calibration can consider that the forecasters' behavior might be different for different truth values t_n. In contrast, LLO calibration assumes symmetric beta distributions for $t_n = 0$ and $t_n = 1$ and therefore symmetric forecasting behavior for true and false queries, which might not hold for real forecasts. A more detailed comparison of beta calibration and LLO calibration including example calibration functions can be found in Section 3.5.

2.4. Hierarchical Symmetric Beta Fusion Model

Since modeling the forecasts with symmetric beta distributions results in calibrating them with the LLO calibration function (Section 2.3), we are interested in comparing the original beta fusion model using asymmetric beta distributions to a symmetric beta fusion model using symmetric beta distributions, which assumes humans to show symmetric forecasting behavior given true or false queries. In the Hierarchical Symmetric Beta Fusion Model, we thus model forecasts x_n^k with two symmetric beta distributions with parameters $\mu_0^k = \mu^k$, $\rho_0^k = \rho^k$ and $\mu_1^k = 1 - \mu^k$, $\rho_1^k = \rho^k$. We set a beta prior on μ^k with mean u^μ and maximum variance proportion p^μ and model ρ^k with a beta distribution with mean u^ρ and maximum variance proportion p^ρ. Similar to the asymmetric beta fusion models, the proportion π of true queries is modeled with an uninformed prior Beta(1,1). Figure 3(a) shows the graphical model of the Hierarchical Symmetric Beta Fusion Model. All modeling distributions and priors are given as

$$\begin{aligned} t_n &\sim \mathrm{Bernoulli}(\pi) & \pi &\sim \mathrm{Beta}(1,1) \\ x_n^k | t_n = 0 &\sim \mathrm{Beta}'(\mu^k, \rho^k) & u^\mu &\sim \mathrm{Beta}(1,1) \\ x_n^k | t_n = 1 &\sim \mathrm{Beta}'(1-\mu^k, \rho^k) & p^\mu &\sim \mathrm{Beta}(1,1) \\ \mu^k &\sim \mathrm{Beta}'(u^\mu, p^\mu) & u^\rho &\sim \mathrm{Beta}(1,1) \\ \rho^k &\sim \mathrm{Beta}'(u^\rho, p^\rho) & p^\rho &\sim \mathrm{Beta}(1,1). \end{aligned} \tag{2.9}$$

The model parameters μ^k, ρ^k, and π are estimated from labeled training data using Gibbs sampling. For fusing unseen forecasts x_n^k, the posterior probability of their true label t_n can be computed



(a) Hierarchical Symmetric Beta Fusion Model (b) Non-Hierarchical Symmetric Beta Fusion Model

Figure 3. *Graphical models of the hierarchical (a) and non-hierarchical (b) symmetric beta fusion models.*

analytically given the learned model parameters:

$$p(t_n = 0|\boldsymbol{x_n}, \boldsymbol{\mu}, \boldsymbol{\rho}, \pi) \propto (1 - \pi) \prod_{k=1}^{K} \text{Beta}'(x_n^k; \mu^k, \rho^k), \tag{2.10}$$

$$p(t_n = 1|\boldsymbol{x_n}, \boldsymbol{\mu}, \boldsymbol{\rho}, \pi) \propto \pi \prod_{k=1}^{K} \text{Beta}'(x_n^k; 1 - \mu^k, \rho^k). \tag{2.11}$$

2.5. Non-hierarchical Symmetric Beta Fusion Model

As for the asymmetric beta fusion model, we also examine a non-hierarchical version of the symmetric beta fusion model. In the non-hierarchical Symmetric Beta Fusion Model, all forecasters' forecasts x_n^k are modeled with the same two symmetric beta distributions for $t_n = 0$ and $t_n = 1$ with parameters $\mu_0 = \mu$, $\rho_0 = \rho$ and $\mu_1 = 1 - \mu$, $\rho_1 = \rho$. The priors for μ and ρ are uniform distributions Beta(1,1). As in all models, we set an uninformative prior Beta(1,1) on the proportion of true queries π. The graphical model of the non-hierarchical Symmetric Beta Fusion Model is presented in Figure 3(b). An overview of all modeling distributions and priors is given as

$$\begin{aligned} t_n &\sim \text{Bernoulli}(\pi) & \mu &\sim \text{Beta}(1,1) \\ x_n^k|t_n = 0 &\sim \text{Beta}'(\mu, \rho) & \rho &\sim \text{Beta}(1,1) \\ x_n^k|t_n = 1 &\sim \text{Beta}'(1 - \mu, \rho) & \pi &\sim \text{Beta}(1,1). \end{aligned} \tag{2.12}$$

The model parameters μ, ρ, and π are estimated from labeled training data using Gibbs sampling. For fusing unseen forecasts x_n^k, the posterior probability of their true label t_n can be computed analytically using equations (2.10) and (2.11) with dropping index k for parameters μ, ρ.

3. Evaluation

We evaluated the four proposed Bayesian models on two data sets consisting of forecasts provided by human subjects, the Turner data set and a new data set that we collected (Section 3.1). Using leave-one-out cross-validation (Section 3.2), we compare their Brier scores, 0–1 losses, and mean absolute errors as performance measures (Section 3.3). In addition, we also compared the performances of the proposed fusion models to the fusion models presented by Turner et al. (2014). For implementing them,

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

we adopted the JAGS code they provided for their models and all specifications given with respect to sampling, e.g., the number of samples and initial values.

3.1. Data sets

We evaluated the performances of the proposed Bayesian fusion models and the reference models by Turner et al. (2014) on two data sets, namely the Turner data set (Section 3.1.1), which is the data set Turner et al. (2014) used for evaluating their fusion models, and our new data set (Section 3.1.2).

3.1.1. Turner data set

The Turner data set[1] is a subset of the The Good Judgment data set,[2] containing 176 geopolitical statements in the form of ‘Will event X happen by date Y?’, e.g., ‘There will be a military coup in Venezuela in 2011’. All statements are binary, so they are either true or false, but at the time of data collection, all events were unresolved, so their outcome could not be known yet. After completion of the study, 37 of 176 statements turned out to be true, while the remaining 139 statements resolved as false.

Human subjects could reply to the given items through a web page. They provided their estimate of the probability that the respective statement will resolve to true for as many statements as they wanted. The provided probabilities are between 0 and 1, accurate to 2 decimal places. To avoid problems with estimates of exactly 0 or 1, we preprocessed these estimates and changed them to 0.001 and 0.999, respectively, as also done in the work of Turner et al. (2014).[3] 1,290 subjects provided a total number of 11,551 probability estimates. The maximum number of replies per subject is 127; on average, a subject provided 8.25 probability estimates.

3.1.2. Knowledge Test Confidence data set

In this work, we publish a new data set called the Knowledge Test Confidence (KTeC) data set.[4] It consists of the confidence judgments of 85 forecasters to 180 knowledge statements of which 90 statements are true and 90 statements are false. There are easy statements, e.g., ‘Elephants are mammals’, and hard statements, at least for our participant pool, e.g., ‘Port Moresby is the capital of Papua New Guinea’.

The data were collected in a probabilistic modeling class at the University of Osnabrück and were part of the lessons on proper scoring rules. The students attending this class were asked to generate statements that are easy to understand, do not contain negations, and cover the whole range from easy to hard statements. They were told that for an easy query 80%–90% of their peers should know the statement’s truth value, for hard queries only 60%–70%. Most of the resulting statements test general knowledge, some are specific to student life in Osnabrück, and some were deliberately designed as trick questions (e.g., ‘The official language of the United States is English’). The students who provided the statements and other students in a couple of following years voluntarily and completely anonymously provided their confidence on the truth of each statement through an online questionnaire. A confidence of 0 indicates that a subject is convinced that the statement is wrong, whereas a confidence of 1 indicates a strong belief that the statement is correct. Subjects could provide their confidences not on a continuous scale, but in 11 steps of 0.1. As for the Turner data set, for the following evaluations we again preprocessed 0 to 0.001 and 1 to 0.999.[5]

85 students provided a total number of 15,300 probability estimates. Thus, each subject replied to all 180 statements.

[1] https://webfiles.uci.edu/msteyver/codeanddata/forecastingdata.csv.

[2] https://dataverse.harvard.edu/dataverse/gjp.

[3] We also evaluated different score corrections, namely (0.01,0.99) and (0.025,0.975), which did not change the results significantly.

[4] The data set is provided in the supplementary material and at https://osf.io/ae25w/.

[5] We also evaluated different score corrections, namely (0.01,0.99) and (0.025,0.975), which did not change the results significantly.

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

3.2. Cross-validation

In order to evaluate the different models on the data sets described above (Section 3.1), we split the data into training and test sets using leave-one-out (LOO) cross-validation. While Turner et al. (2014) evaluated with 10-fold cross-validation, we preferred leave-one-out cross-validation over k-fold cross-validation since it allows training the model on almost the entire data set, which reduces the evaluation bias. Also, it comes with zero randomness in the partitioning of the data and is therefore straightforwardly reproducible.

If a data set consists of M queries, we obtain M LOO training sets, each including $M-1$ data points. For each of the resulting M training sets, we inferred the posterior distribution of each model's parameters using Gibbs sampling. We implement Gibbs sampling for inference using JAGS (Plummer, 2003). For fitting the model parameters given labeled training data, we ran 2 parallel chains, each consisting of 1,000 samples with a burn-in of 1,000 samples.

For fusing the one example in the test set, one could now use the means of the obtained posterior distributions as point estimates for the parameters and compute the posterior over the true label t_n analytically given these point estimates. However, in order to consider the uncertainty of our parameter estimates, we computed the posterior over t_n analytically for each sample of the model parameters' posterior distribution, as for example in (2.4), and averaged all obtained posteriors to the final fused forecast.

3.3. Performance measures

As performance measures, we consider Brier score, 0–1 loss, and mean absolute error. The Brier score (Brier, 1950) is a popular metric for quantifying human forecast performance, used by e.g., Karvetski et al. (2013), Turner et al. (2014), Hanea et al. (2021), and Satopää (2022). It is a strictly proper scoring rule (Murphy, 1973), meaning that it is optimized when people report their true beliefs of the probability instead of intentionally providing more or less extreme probabilities. The Brier score is defined as the mean squared error between the predicted probabilities x_n and the true labels t_n,

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}(x_n - t_n)^2. \tag{3.1}$$

Thus, the best attainable Brier score is 0 and the worst is 1. Interestingly, if a forecaster always provides 0.5 as her estimate, the resulting Brier score will be 0.25. Thus, a model should at least achieve a Brier score below 0.25.

It is controversial whether the Brier score is a suitable metric for comparing the performances of different forecasting systems, since as a strictly proper scoring rule it was originally developed in order to measure if forecasters report their true beliefs, not to compare different forecasters (Steyvers et al., 2014). Also, it can be dominated by outliers (Canbek et al., 2022), though not as much as e.g., the log-loss. Still, it is commonly used for comparing forecasters' performances (Baron et al., 2014; Hanea et al., 2021; Karvetski et al., 2013; Ranjan & Gneiting, 2010; Satopää, 2022; Turner et al., 2014), so we report it here, too.

The 0–1 loss describes the proportion of incorrect forecasts to the total number of all forecasts,

$$\mathrm{L}_{01} = \frac{1}{N}\sum_{n=1}^{N}\begin{cases} 1 & \text{if} \quad |t_n - x_n| \geq 0.5 \\ 0 & \text{else} \end{cases}, \tag{3.2}$$

and thus ranges between 0 and 1, with lower values indicating better performances. In comparison to the Brier score, the 0–1 loss is more easily interpretable and directly compares the forecasters' performances. However, it disregards their uncertainty by considering a forecast as correct, if its corresponding probability is closer to the true label t_n of the respective query.

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

To overcome the limitations of the Brier score and 0–1 loss, we additionally evaluate the different fusion methods in terms of mean absolute error (Canbek et al., 2022; Ferri et al., 2009). Mean absolute error (MAE) measures the absolute difference between the forecasted probability and the true label:

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N} |x_n - t_n|. \tag{3.3}$$

Similar to the Brier score (mean squared error) and 0–1 loss, it ranges between 0 and 1, with lower values indicating higher performance. However, in contrast to the Brier score, which can be dominated by outliers, MAE is more robust to outliers (Canbek et al., 2022). Also, it is straightforwardly interpretable and a more natural and intuitive metric for comparing different fusion models without disregarding their uncertainty. However, note that MAE is an improper scoring rule (Buja et al., 2005), so it incentivizes overconfident forecasts.

3.4. *Model performances*

On both data sets described in Section 3.1, we evaluate the four Bayesian fusion models introduced in Section 2 in terms of Brier score, 0–1 loss, and mean absolute error (MAE). In addition, we compare our fusion methods' performances to the models by Turner et al. (2014). Turner et al. (2014) present several Bayesian fusion models, which explicitly consider the calibration of forecasts with the LLO calibration function. Their key question is whether it is better first to average the forecasts or first to calibrate them. Hence, the proposed models are three non-hierarchical models, Average then Calibrate (ATC), Calibrate then Average (CTA), Calibrate then Average using log-odds (CTALO), and two hierarchical models, Hierarchical Calibrate then Average (HCTA), and Hierarchical Calibrate then Average on log-odds (HCTALO).[6] For reference, they also evaluate the performance of Unweighted Linear Opinion Pool (ULINOP) as a baseline. Since ULINOP is known to be biased toward 0.5 (Baron et al., 2014), we additionally evaluate the performance of Probit Average (PAVG) (Satopää et al., 2023) as another benchmark. For PAVG, we first transform all forecasts with probit, then average the transformed forecasts, and finally transform this average back to probability score. Here, we investigate how a normative model that implicitly calibrates the forecasts by modeling them with beta distributions will perform relative to all these models. In particular, we investigate whether the normative approach increases the performance of the fused forecast.

Figure 4 shows the means and standard errors of the mean of Brier score, 0–1 loss, and MAE on the Turner data set (Section 3.1.1). The beta fusion models introduced in this work are abbreviated as HB (Hierarchical Beta Fusion Model), B (non-hierarchical Beta Fusion Model), HSB (Hierarchical Symmetric Beta Fusion Model), and SB (non-hierarchical Symmetric Beta Fusion Model). According to the Brier score, the three non-hierarchical Turner models ATC, CTA, and CTALO perform best with BS $\approx$ 0.125. These results are different to the results reported by Turner et al. (2014), which were obtained using 10-fold cross-validation and favored the HCTALO model. The best beta model is HSB with BS = 0.151, which performs comparably to HB and the hierarchical Turner models HCTA and HCTALO. The non-hierarchical beta models B and SB perform worst with the Brier scores of about 0.25, which is close to the performance of a forecaster that always forecasts 0.5. As per 0–1 loss, the ranking of the models is different. HSB performs similarly to ATC, CTA, CTALO, and HCTA with $\mathrm{L}_{01} = 0.176$, HB is approaching ($\mathrm{L}_{01} = 0.21$). The hierarchical Turner model HCTALO ($\mathrm{L}_{01} = 0.267$) performs clearly worse than both hierarchical beta models. According to MAE, both hierarchical beta models HB (MAE = 0.219) and HSB (MAE = 0.185) perform best. All non-hierarchical models, Turner and beta models, perform similarly (MAE $\approx$ 0.25), while the hierarchical Turner models perform worst, similarly to ULINOP and PAVG with MAE $\approx$ 0.4. Consistently over all performance

[6] In our evaluations, we used the JAGS code provided in the work of Turner et al. (2014). However, note that in their work the implementation of HCTALO is significantly different from the implementation of HCTA.

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

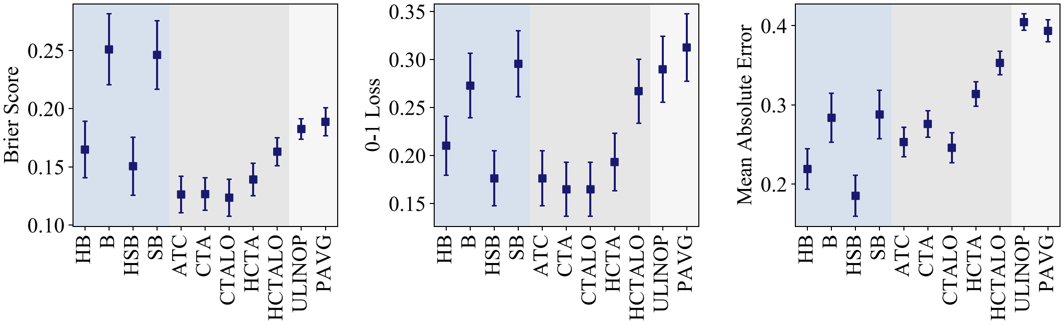


Figure 4. *Model performances on the Turner data set according to Brier score, 0–1 loss, and mean absolute error. We compare the scores' means and standard errors of the mean of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average then Calibrate (ATC), Calibrate then Average (CTA), Calibrate then Average using log-odds (CTALO), Hierarchical Calibrate then Average (HCTA), and Hierarchical Calibrate then Average on log-odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).*

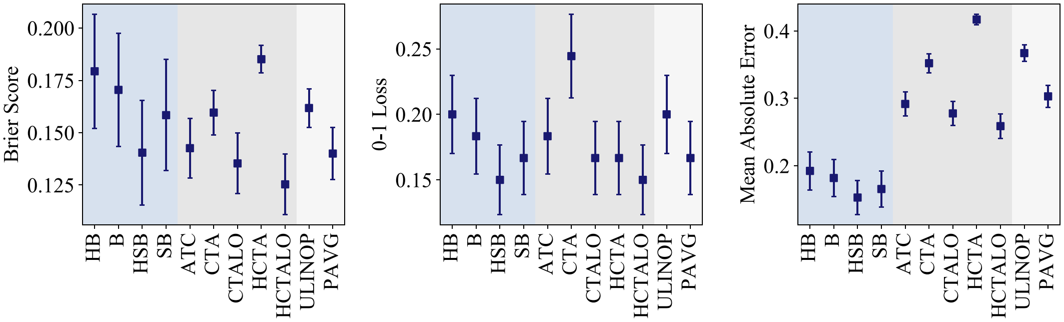


Figure 5. *Model performances on the Knowledge Test Confidence data set according to Brier score, 0–1 loss, and mean absolute error. We compare the scores' means and standard errors of the mean of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average then Calibrate (ATC), Calibrate then Average (CTA), Calibrate then Average using log-odds (CTALO), Hierarchical Calibrate then Average (HCTA), and Hierarchical Calibrate then Average on log-odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).*

measures, HSB outperforms HB. Also, the hierarchical beta models outperform the non-hierarchical beta models, while the non-hierarchical Turner models outperform the hierarchical Turner models.

In Figure 5, we compare the performances of beta and Turner fusion models on the newly introduced KTeC data set. Compared to the results on the Turner data set, the differences between the models' performances are generally smaller over all three performance measures Brier score, 0–1 loss, and MAE. Based on the Brier score, HCTALO performs best with BS = 0.125, however, CTALO, ATC, HSB, and PAVG perform quite similarly. HB and HCTA perform worst with Brier scores of BS = 0.179 and BS = 0.185. However, as per 0–1 loss, CTA performs worst, and HSB and HCTALO are performing best with L_{01} = 0.156 and L_{01} = 0.15. According to MAE, all beta fusion models clearly outperform the Turner models and perform quite comparably. Still, HSB is again the best performing beta fusion model with an MAE of 0.153.

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

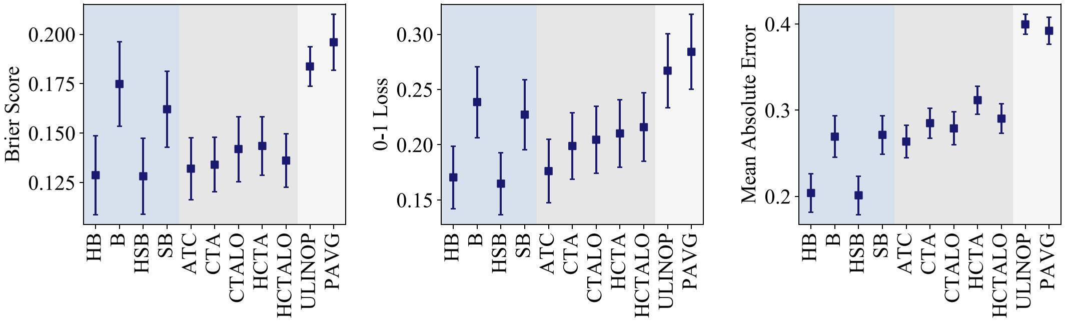


Figure 6. *Model performances on the reduced Turner data set consisting of a subset of the 20 forecasters of the Turner data set that provided the most forecasts. We compare the means and standard errors of the mean of Brier scores, 0–1 losses, and mean absolute errors of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average then Calibrate (ATC), Calibrate then Average (CTA), Calibrate then Average using log-odds (CTALO), Hierarchical Calibrate then Average (HCTA), and Hierarchical Calibrate then Average on log-odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).*

In the evaluations shown so far, we combine the forecasts of 1,290 forecasters for the Turner data set and 85 forecasters for the KTeC data set. For both data sets, the Hierarchical Symmetric Beta Fusion model (HSB) is best or among the best models according to 0–1 loss or MAE. However, according to Brier score, some Turner models outperform HSB on both data sets, which might indicate that HSB or the beta models in general are overconfident. To investigate this, we also consider two subsets of the two data sets that contain fewer forecasters since fusing a lower number of forecasters should attenuate overconfidence. For the Turner data set, the subset of forecasters must be chosen with care, since all forecasters only provided forecasts for only a subset of queries. To be able to evaluate fusion methods, we must guarantee that all queries are answered by at least two forecasters, which we can then fuse. In order to do so, we selected the 20 forecasters providing the most forecasts. The results on the reduced Turner data set are shown in Figure 6. As for the full Turner data set, the hierarchical beta fusion models outperform the non-hierarchical ones, achieve a 0–1 loss ($L_{01} = 0.171$ for HB and $L_{01} = 0.165$ for HSB) comparable to the best Turner model ATC ($L_{01} = 0.176$), and outperform all other models clearly according to MAE with MAE ≈ 0.2. In addition, and in contrast to the full Turner data set, here HB (BS = 0.129) and HSB (BS = 0.128) also perform comparably to the best Turner model ATC (BS = 0.132) regarding Brier score. Thus, for the reduced Turner data set, the hierarchical beta fusion models HB and HSB are among the best fusion models for all performance measures.

For the KTeC data set, it is more straightforward to create a reduced data set since all forecasters replied to all queries. Therefore, we simply selected the first 10 forecasters as a subset. In Figure 7, we see that similar to the reduced Turner data set, on the reduced KTeC data set, HSB is the best model according to all three performance measures ($\mathrm{BS} = 0.124, L_{01} = 0.15, \mathrm{MAE} = 0.192$). As per Brier score and 0–1 loss, HCTALO ($\mathrm{BS} = 0.132, L_{01} = 0.156$) performs comparably, but regarding MAE, again all beta fusion models clearly outperform all Turner models. Similar to the full KTeC data set, the symmetric beta model (HSB) generally outperforms the asymmetric one (HB) and the hierarchical beta models achieve better scores than the non-hierarchical ones.

3.5. Beta calibration vs LLO

The results presented in Section 3.4 show that the Hierarchical Symmetric Beta Fusion Model (HSB) outperforms the Hierarchical Beta Fusion Model (HB). While this outcome is rather unexpected, since

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

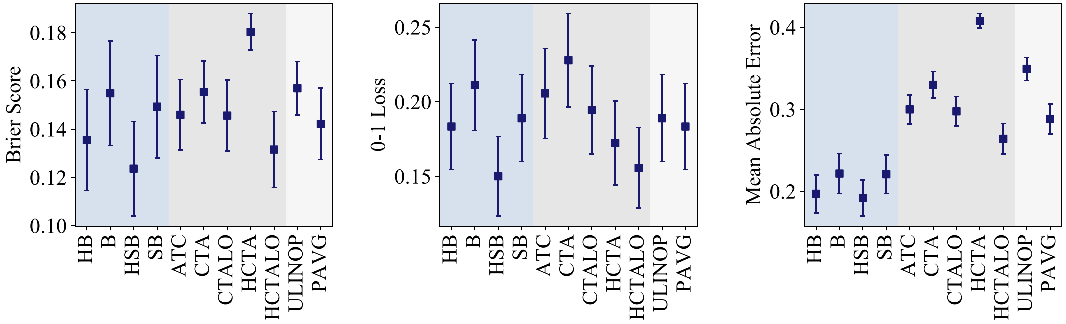


Figure 7. *Model performances on the reduced Knowledge Test Confidence (KTeC) data set consisting of a subset of the first 10 forecasters of KTeC data set. We compare the means and standard errors of the mean of Brier scores, 0–1 losses, and mean absolute errors of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average then Calibrate (ATC), Calibrate then Average (CTA), Calibrate then Average using log-odds (CTALO), Hierarchical Calibrate then Average (HCTA), and Hierarchical Calibrate then Average on log-odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).*

HSB constrains the modeling beta distributions to be symmetric and is therefore less expressive than HB, we can explain it with the calibration functions that are implied by modeling the forecasts with symmetric or asymmetric beta distributions. As shown in Section 2.3, by modeling the forecasts with beta distributions conditioned on the true label, we implicitly calibrate them using the beta calibration function (2.7). If the beta distributions are symmetric, the beta calibration function reduces to the LLO calibration function (2.8).

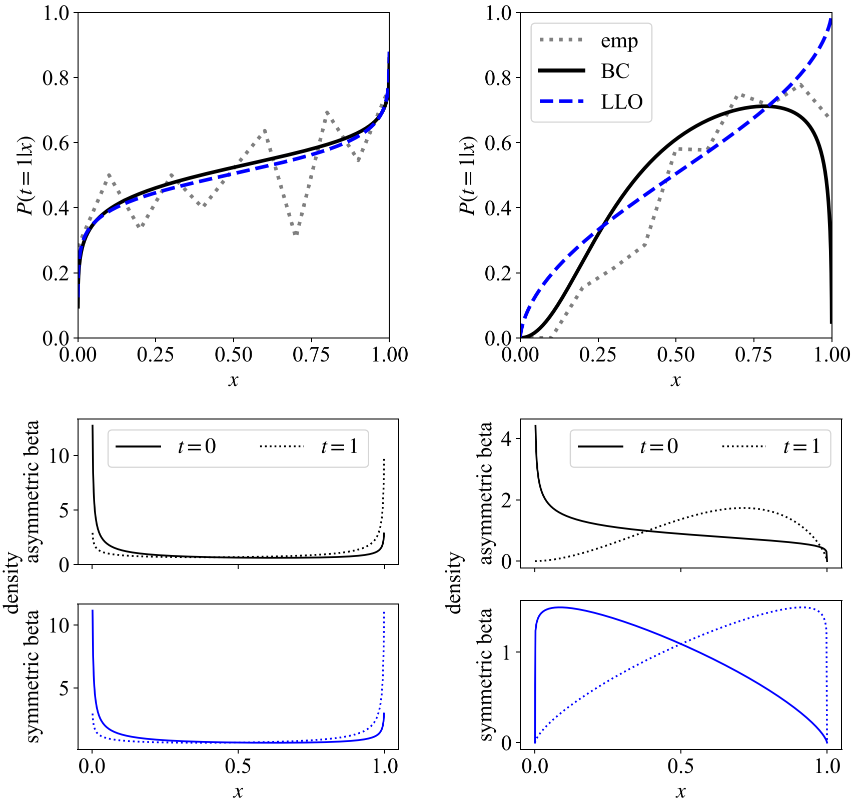
In most cases, the LLO and beta calibration function do not differ significantly. However, there are special cases, in which the beta calibration deviates drastically from the LLO calibration function. From (2.8), we can directly see that $\mathrm{LLO}(0) = 0$ and $\mathrm{LLO}(1) = 1$ if $\gamma = \beta - \alpha > 0$ with $\alpha = \alpha_0^k = \beta_1^k$ and $\beta = \beta_0^k = \alpha_1^k$. The latter condition should hold for the most forecasters, since otherwise they would be biased toward always predicting the wrong answer.

In contrast, also for such unbiased forecasters, for which $\alpha_0^k < \beta_0^k$ and $\alpha_1^k > \beta_1^k$, the beta calibration function $\mathrm{BC}(x)$ is not always defined at $x = 0$ and $x = 1$, depending on the beta parameters. In particular, looking at (2.7), we see that

$$\begin{aligned} \lim_{x\to 0} \mathrm{BC}(x) &= 1 \quad \text{if} \quad \alpha_1^k - \alpha_0^k < 0 \quad \text{and} \\ \lim_{x\to 1} \mathrm{BC}(x) &= 0 \quad \text{if} \quad \beta_0^k - \beta_1^k < 0. \end{aligned} \tag{3.4}$$

In Figure 8, we show the calibration curves of two exemplary forecasters from the KTeC data set (top row) together with the densities of the respective beta distributions for $t = 0$ and $t = 1$ when assuming asymmetric or symmetric beta distributions (bottom row). The respective parameters of the beta distributions that model their forecasts and define their calibration curves are taken from training split 1 of the cross-validation done for HB and HSB described in Section 3.2. Figure 8(a) shows the calibration curves and corresponding beta distributions of forecaster 57 with parameters $\alpha_0^{57} = 0.41, \beta_0^{57} = 0.65, \alpha_1^{57} = 0.67, \beta_1^{57} = 0.47$ for the asymmetric Hierarchical Beta Fusion Model (HB) or the beta calibration function, respectively, and parameters $\alpha_0^{57} = \beta_1^{57} = 0.44, \beta_0^{57} = \alpha_1^{57} = 0.65$ for the Hierarchical Symmetric Beta Fusion Model (HSB) and the LLO calibration function. We can see that the learned beta distributions for HB (asymmetric) and HSB (symmetric) look very similar. Also, LLO and beta calibration curves look very similar since $\alpha_1^{57} - \alpha_0^{57} > 0$ and $\beta_0^{57} - \beta_1^{57} > 0$.

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press



(a) BC and LLO similar

(b) BC and LLO different

Figure 8. *The empirical, beta calibration (BC), and LLO curves of two exemplary forecasters from the Knowledge Test Confidence data set (top row) together with the respective asymmetric and symmetric beta distributions (bottom row). (a) shows the calibration curves and respective beta distributions of forecaster 57, where LLO and beta calibration curves are similar. (b) shows the calibration curves and beta distributions of forecaster 46, for which the beta calibration function tends to 0 for $x \longrightarrow 1$ causing miscalibrations.*

In contrast, in Figure 8(b), we see that for forecaster 46, the beta calibration curve looks very different from the LLO curve. Since the forecaster is modeled with parameters $\alpha_0^{46} = 0.73$, $\beta_0^{46} = 1.16$, $\alpha_1^{46} = 2.81$, $\beta_1^{46} = 1.73$ for HB or beta calibration and $\alpha_0^{46} = \beta_1^{46} = 1.07, \beta_0^{46} = \alpha_1^{46} = 1.72$ for HSB or LLO, $\beta_0^{46} - \beta_1^{46} < 0$ and BC(x) tends to 0 for $x \longrightarrow 1$. This can also be seen in the corresponding beta distributions in the bottom of Figure 8(b). If we assume asymmetric beta distributions, at $x = 1$ the probability density for $t = 0$ is higher than the probability density for $t = 1$, leading to a beta calibration function that tends to 0 for $x \longrightarrow 1$. If we assume symmetric beta distributions, this does not happen. In this case, fusing with HB and thereby calibrating with beta calibration can induce miscalibration of forecasts close to 1 that lead to forecasting the opposite of the forecast that was originally provided. This can result in worse performance of the HB fusion model in comparison to the HSB fusion model, which we see in the results presented in Section 3.4.

4. Discussion and conclusion

In this work, we presented a family of normative generative models for fusing probability forecasts. Since uncertainty over probabilities is commonly modeled with the beta distribution, in our normative fusion models, we model each forecaster's probability forecasts with beta distributions conditioned on their true label. We compare different variants of this model including hierarchical and non-hierarchical as well as asymmetric and symmetric beta fusion models. Given the respective model, new unseen probability estimates can be fused by inferring their true label. The obtained fused forecast is Bayes optimal given the model's assumptions. While previous approaches explicitly calibrate the considered forecasts using the LLO calibration function (Lee & Danileiko, 2014; Turner et al., 2014), the proposed beta fusion models implicitly calibrate the probability estimates provided by the forecasters with the beta calibration function, which accommodates the LLO calibration function as a special case.

We evaluated the proposed models on a data set by Turner et al. (2014) and the newly introduced KTeC data set, also including two smaller subsets of these two data sets. In this vein, we also compared the proposed beta fusion models to the models by Turner et al. (2014), which fuse forecasts by averaging and calibrating them using the LLO calibration function. Looking at the results of all four data sets, i.e., the two full data sets and their respective reduced subsets, we can observe some general findings.

The hierarchical beta models generally outperform the non-hierarchical beta models. This is expected behavior because the hierarchical models are able to model each individual forecaster's behavior, which can be different, while the non-hierarchical models assume exchangeable forecasters, model all forecasters' collective behavior, and therefore discard valuable information. However, the Turner and KTeC data sets differ in the magnitude of the difference between the performances of non-hierarchical and hierarchical beta fusion models. For the Turner data set, this difference is greater than for the KTeC data set, since in the Turner data set the beta parameters of different forecasters are more variable than in the KTeC data set. Therefore, modeling all forecasters in the Turner data set with the same beta parameters causes comparably inferior performance.

Among the beta fusion models, the Hierarchical Symmetric Beta Fusion Model (HSB) shows the best performance. In particular, it outperforms the (asymmetric) Hierarchical Beta Fusion Model (HB), although HB does not constrain the modeling beta distributions to be symmetric and is therefore more expressive than HSB. As we discussed in Section 3.5, the reason for this is the calibration functions implied by the beta distributions, beta calibration for HB and LLO for HSB. Depending on the parameters learned for HB, beta calibration can lead to miscalibration of forecasts which causes worse performance of HB compared to HSB. Therefore, and in line with the results presented in Section 3.4, we recommend using the Hierarchical Symmetric Beta Fusion Model (HSB) instead of the Hierarchical Beta Fusion Model, and LLO calibration instead of beta calibration accordingly.

The conclusions discussed above are all consistent for the three performance measures Brier score, 0–1 loss, and MAE. However, as we report in Section 3.4, different measures suggest different models as the best-performing model. As we mentioned in Section 3.1, it is not clear which measure should be preferred. Brier score is commonly used (Baron et al., 2014; Hanea et al., 2021; Karvetski et al., 2013; Ranjan & Gneiting, 2010; Satopää, 2022; Turner et al., 2014), but criticized for not being appropriate for comparing forecasters, because it was developed for measuring whether forecasters report their true beliefs (Steyvers et al., 2014). Also, Brier score can be dominated by outliers (Canbek et al., 2022). On the other hand, the 0–1 loss directly compares the forecasters' performances and is easily interpretable but disregards their uncertainty. MAE considers the forecasters' uncertainty, is straightforwardly interpretable, and is robust to outliers (Canbek et al., 2022). However, it is an improper scoring rule (Buja et al., 2005) and incentivizes overconfident forecasts.

Since Brier score, 0–1 loss, and MAE have different strengths and weaknesses, we reported all three measures in our work. Interestingly, the differences between the results according to different measures reveal something about the models' properties. The hierarchical beta fusion models always outperform all other models regarding MAE. Thus, at least according to this measure, the normative models outperform the models proposed by Turner et al. (2014). On the two full data sets, some Turner models achieve lower, i.e., better Brier scores than the HSB, but these are different models depending on the respective data set, ACT, CTA, CTALO, and HCTA on the Turner data set, HCTALO on the KTeC data set. Still, according to 0–1 loss, HSB is always competing with these best Turner models

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

and outperforms them in terms of MAE. On the reduced data sets, also HSB's Brier score is better or competing to all other models.

The main reason why the hierarchical beta models (HB and HSB) achieve better MAE scores but worse Brier scores than some of Turner's models on the full data sets is their overconfidence. This overconfidence is a direct result of the conditional independence assumption of our beta fusion models (given the label, each subject provides independent probabilistic forecasts). This assumption is not met in the data. Since the forecasters respond to the same questions and share information and knowledge, their forecasts are not independent but tend to increase when other forecasters' forecasts increase. For example, the forecasters in our participant pool for the KTeC data set did not know the capital of Papua New Guinea but knew the answers to questions related to their university. Since they were students from Osnabrück in Germany, they shared knowledge about their university but consistently had little geographic knowledge about Papua New Guinea. If forecasts are combined assuming conditional independence, the fused forecast's uncertainty is usually reduced. This effect becomes stronger when more forecasts are fused. Unfortunately, if forecasts that are not actually independent are fused in this way, the fused forecast can be more confident than it should be (Trick & Rothkopf, 2022). Those overconfident forecasts can lead to high Brier scores, because the Brier score drastically punishes wrong forecasts with high confidence. If fewer forecasts are combined, the fused forecast is less overconfident, which is why the hierarchical beta models achieve better Brier scores than Turner's models on the two reduced data sets. The beta fusion models' overconfidence is also the reason why the standard errors of the mean, shown as error bars in the figures, are larger for the beta fusion models than for Turner's models regarding Brier score and MAE. More confident fusion models lead to higher variability in the Brier and MAE scores of different splits of LOO cross-validation and thus to higher standard errors in these measures.

Since the forecasts of multiple human experts are rarely independent (Jacobs, 1995; Wiper & French, 1995), future work on combining forecasts should include the possibility to model correlations between forecasters to take into account the shared questions and knowledge of the forecasters, which will further increase the performance of the fused forecast. This could be realized, for example using correlated beta distributions (Arnold & Ghosh, 2017; Moschen & Carvalho, 2023; Trick et al., 2023).

Supplementary material. The supplementary material including the new Knowledge Test Confidence data set can be found at https://doi.org/10.1017/jdm.2023.39.

Data availability statement. The new Knowledge Test Confidence data set can be found in the supplementary material at https://doi.org/10.1017/jdm.2023.39. In addition, the data set is available at https://osf.io/ae25w/.

Funding statement. This work was supported by the German Federal Ministry of Education and Research (BMBF) in the project IKIDA (Grant No. 01IS20045).

Competing interest. The authors have no competing interests to declare.

References

Arnold, B. C., & Ghosh, I. (2017). Bivariate beta and Kumaraswamy models developed using the Arnold–Ng bivariate beta distribution. *REVSTAT—Statistical Journal*, *15*(2), 223–250.

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

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. London, Springer.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219.

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

Britten, G. L., Mohajerani, Y., Primeau, L., Aydin, M., Garcia, C., Wang, W.-L., Pasquier, B., Cael, B. B., & Primeau, F. W. (2021). Evaluating the benefits of Bayesian hierarchical methods for analyzing heterogeneous environmental data sets: A case study of marine organic carbon fluxes. *Frontiers in Environmental Science*, *9*, 491636.

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

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398.

Buja, A., Stuetzle, W., & Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working Draft*. https://sites.stat.washington.edu/wxs/Learning-papers/paper-proper-scoring.pdf

Canbek, G., Temizel, T. T., & Sagiroglu, S. (2022). PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. *SN Computer Science*, *4*(1), 13.

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

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, *4*(1), 39–46.

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

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford: Oxford University Press.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In *The Oxford handbook of probability and philosophy* (pp. 519–542). Oxford: Oxford University Press.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38.

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

Graefe, A. (2018). Predicting elections: Experts, polls, and fundamentals. *Judgment and Decision Making*, *13*(4), 334–344.

Graham, J. R. (1996). Is a group of economists better than one? Than none? *Journal of Business*, *69*(2), 193–232.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl Jr. , K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110–1130.

Hanea, A., Wilkinson, D. P., McBride, M., Lyon, A., van Ravenzwaaij, D., Singleton Thorn, F., Gray, C., Mandel, D. R., Willcox, A., Gould, E., Smith, E. T., Mody, F., Bush, M., Fidler, F., Fraser, H., & Wintle, B. C. (2021). Mathematically aggregating experts' predictions of possible futures. *PLoS One*, *16*(9), 1–24.

Jacobs, R. A. (1995). Methods for combining experts' probability assessments. *Neural Computation*, *7*(5), 867–888.

Ji, D., Smyth, P., & Steyvers, M. (2020). Can I trust my fairness metric? Assessing fairness with unlabeled data and Bayesian inference. *Advances in Neural Information Processing Systems*, *33*, 18600–18612.

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, *10*(4), 305–326.

Kull, M., Silva Filho, T., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics* (pp. 623–631). PMLR.

Lee, M. D. (2018a). Bayesian methods in cognitive modeling. In *The Stevens' handbook of experimental psychology and cognitive neuroscience* (vol. 5, pp. 37–84). Hoboken, NJ: Wiley.

Lee, M. D. (2018b). In vivo: Multiple approaches to hierarchical modeling. In S. Farrell & S. Lewandowsky (Eds.), *Computational modeling of cognition and behavior*. Cambridge: Cambridge University Press. https://webfiles.uci.edu/mdlee/LeeInVivo.pdf?uniq=fe8jrx

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 258–272.

Lichtendahl Jr., K. C., Grushka-Cockayne, Y., Jose, V. R., & Winkler, R. L. (2022). Extremizing and antiextremizing in Bayesian ensembles of binary-event forecasts. *Operations Research*, *70*(5), 2998–3014.

Lindley, D. V. (1985). Reconciliation of discrete probability distributions. *Bayesian Statistics*, *2*, 375–390.

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

Minson, J. A., Mueller, J. S., & Larrick, R. P. (2018). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science*, *64*(9), 4177–4192.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*(20), 7176–7184.

Moschen, L. M., & Carvalho, L. M. (2023). Bivariate beta distribution: Parameter inference and diagnostics. *Preprint*, arXiv:2303.01271.

Murphy, A. H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology and Climatology*, *12*(1), 215–223.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, Vienna, Austria.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(1), 71–91.

https://doi.org/10.1017/jdm.2023.39 Published online by Cambridge University Press

Satopää, V. A. (2022). Regularized aggregation of one-off probability predictions. *Operations Research*, *70*(6), 3558–3580.

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

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2023). Decomposing the effects of crowd-wisdom aggregators: The bias–information–noise (BIN) model. *International Journal of Forecasting*, *39*(1), 470–485.

Silver, I., Mellers, B. A., & Tetlock, P. E. (2021). Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology*, *96*, 104157.

Steyvers, M., Wallsten, T. S., Merkle, E. C., & Turner, B. M. (2014). Evaluating probabilistic forecasts with Bayesian signal detection models. *Risk Analysis*, *34*(3), 435–452.

Trick, S., & Rothkopf, C. (2022). Bayesian classifier fusion with an explicit model of correlation. In *International conference on artificial intelligence and statistics*. PMLR.

Trick, S., Rothkopf, C. A., & Jäkel, F. (2023). Parameter estimation for a bivariate beta distribution with arbitrary beta marginals and positive correlation. *METRON*, *81*, 163–180.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*(3), 261–289.

Wang, J., Liu, Y., & Chen, Y. (2021). Forecast aggregation via peer prediction. In *Proceedings of the AAAI conference on human computation and crowdsourcing*.

Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, *33*(1), 325–336.

Wiper, M. P., & French, S. (1995). Combining experts' opinions using a normal-Wishart model. *Journal of Forecasting*, *14*(1), 25–34.

Cite this article: Trick, S., Rothkopf, C. A., and Jäkel, F. (2023). A normative model for Bayesian combination of subjective probability estimates. *Judgment and Decision Making*, e40. https://doi.org/10.1017/jdm.2023.39