Article:
Powell, Ben orcid.org/0000-0002-0247-7713, Satopaa, Ville, MacKay, Niall orcid.org/0000-0003-3279-4717 et al. (1 more author) (2024) Skew-Adjusted Extremized-Mean:A Simple Method for Identifying and Learning From Contrarian Minorities in Groups of Forecasters. Decision. 173–193. ISSN 2325-9965

https://doi.org/10.1037/dec0000191

Skew-Adjusted Extremized-Mean: A Simple Method for Identifying and Learning From Contrarian Minorities in Groups of Forecasters

Ben Powell[1*], Ville A. Satopää[2], Niall MacKay[1], and Philip E. Tetlock[3]

[1]Department of Mathematics, University of York, York, UK
[2]Department of Technology and Operations Management, INSEAD, Fontainebleau, France
[3]Department of Psychology, University of Pennsylvania, Philadelphia, USA
[*]Corresponding Author: ben.powell@york.ac.uk

ABSTRACT

Recent work in forecast aggregation has demonstrated that paying attention to contrarian minorities among larger groups of forecasters can improve aggregated probabilistic forecasts. In those papers, the minorities are identified using 'meta-questions' that ask forecasters about their forecasting abilities or those of others. In the current paper, we explain how contrarian minorities can be identified without the meta-questions by inspecting the skewness of the distribution of the forecasts. Inspired by this observation, we introduce a new forecast aggregation tool called *Skew-Adjusted Extremized-Mean* and demonstrate its superior predictive power on a large set of geopolitical and general knowledge forecasting data.

1 Introduction

This paper is motivated by the need to aggregate probabilistic forecasts of geopolitical events. More specifically, it is a response to the work of the Good Judgment Project (GJP) (e.g., Ungar et al. 2012). The GJP was conducted in collaboration with the United States' Intelligence Advanced Research Projects Activity (IARPA) between 2011-15 and centred around an empirical study of the forecasting abilities of a large cohort of individuals. One of the project's core challenges to was to introduce effective methodology for communal or aggregated forecasting.

Aggregation of dispersed information has been at the centre of statistical study since the inception of the discipline (e.g., Laplace 1774, Sheynin 1977, Galton 1907, Bates and Granger 1969, and many others; for comprehensive reviews, see Clemen 1989, Genest et al. 1986, and Zellner et al. 2021). Despite this long history, the topic continues to provide opportunities for new research. One popular direction has sought to identify and correct common biases in the individual forecasts or their simple aggregates such as the (equally-weighted) arithmetic mean. Indeed, the literature has long acknowledged that the aggregated forecasts in certain contexts can

be under-confident (e.g., Lichtenstein et al. 1977). Erev et al. (1994) show how under-confidence can arise from mismatches between forecaster and aggregator behaviour. Ranjan and Gneiting (2010) show both theoretically and empirically that the (weighted) arithmetic mean of forecasts is under-confident and lacks calibration. Satopää (2021) generalize this result to all univariate forecasts and all 'means', defined as aggregators that remain within the (open) convex hull of the individual forecasts.

Such systematic errors offer an opportunity for improvement. In particular, if the aggregate is under-confident, its performance can be improved with a transformation that makes it more confident. In fact, the literature has demonstrated that significant improvements in forecast accuracy can be made when the arithmetic mean of forecasts is *extremized* – that is, adjusted in the direction of the nearest extreme value (Karmarkar, 1978; Lattimore et al., 1992; Baron et al., 2014; Satopää et al., 2014, 2016) or, more precisely, directly away from a prior probability (Satopää et al., 2017; Lichtendahl Jr et al., 2020). The amount of extremization is typically found by choosing the value that improves the chosen aggregator the most on historical forecasting data. Turner et al. (2014) provide a thorough comparison of several extremizing procedures and discuss their provenance in detail.

Another popular direction of research involves 'contrarian minorities'. In particular, Prelec et al. (2017), Martinie et al. (2020), and Palley and Satopää (2021) explain why contrarian minorities should receive more weight in the aggregate forecast. They demonstrate the success of this strategy in aggregating multiple answers to general knowledge (GK) questions, such as 'What is the capital of Pennsylvania?'. Here the contrarian forecasters are identified via 'meta-questions' that ask for forecasters' beliefs about the other forecasters' likely answers. Specifically, the contrarian minority is formed of forecasters who believe that the majority will get the answer wrong. In this sense, the minority is then defined by its contrarianism, which is assumed to indicate better or 'more informed' forecasting.

To the best of our knowledge, these two areas of forecast aggregation have so far remained separate. Motivated by this gap in the literature, the current paper seeks to connect the ideas of 'extremization' and 'contrarian minorities'. Specifically, we explore the following two questions. First, can we identify contrarian minorities based on current forecasts of the event alone and without meta-questions or individual records of historical predictive skill? If so, can we significantly improve the accuracy of common extremization techniques by allowing them to account for contrarian minorities? In the remainder of the paper, we demonstrate that the answer to both of these questions is 'Yes'.

The rest of the paper is organized as follows. Section 2 provides a theoretical motivation for skew-adjustment and explains how it can be applied in aggregation of probability predictions of future events. Section 3 evaluates our skew-adjustment on geopolitical and general knowledge forecasting data and compares its performance against that of several commonly-used forecast aggregators. Section 4 concludes with a discussion of limitations and future research directions of our work.

2 Skew-adjusted Extremization

2.1 Forecast Extremization

Consider a population of forecasts, described by a probability density function $\pi(X)$. Extremization is an ex-post adjustment of an aggregate forecast directly away from a reference value. Even though, in principle, extremization can be applied to any aggregate (Satopää, 2021), in this paper we focus on extremizing the mean forecast $\mu = \mathbb{E}[X]$. If we denote the reference value (e.g., the forecasters' common prior mean) with μ_0, then a parsimonious extremization function is given by

$$\mu_0 + \alpha(\mu - \mu_0), \tag{1}$$

where $\alpha \in [0, \infty)$ determines the amount of extremization. In particular, if $\alpha > 1$, then μ is transformed directly away from the reference value μ_0; else if $\alpha \in [0,1)$, μ is 'anti-extremized' and shrank towards μ_0.

The form (1) is intuitive and highlights the mechanics of extremization. However, given that (1) is nothing but a linear transformation of the mean aggregate μ, it can be expressed equivalently as

$$a + b\mu, \tag{2}$$

where $a = \mu_0(1-\alpha)$ and $b = \alpha$. Form (2) is simpler than form (1). Furthermore, if we have access to a training set of past outcomes and their forecasts, then we can estimate the parameters a and b efficiently with the usual linear regression techniques (e.g., Ravishanker et al. 2002).

2.2 Skewness and Contrarian Minorities

Suppose that the population of forecasts is made up of two subpopulations described by probability density functions $\pi_1(x)$ and $\pi_2(x)$. Assume that the moments of these distributions are well defined and denote the mean and variance of subpopulation $s \in \{1,2\}$ with $\mathbb{E}(X \mid X \sim \pi_s) = \mu_s$ and $\text{var}(X \mid X \sim \pi_s) = \sigma_s^2$. The density function describing the whole population is given by the mixture

$$\pi(x) = w_1\pi_1(x) + w_2\pi_2(x), \tag{3}$$

where $w_1 \geq 0$, $w_2 \geq 0$, and $w_1 + w_2 = 1$.

The weights w_1 and w_2 represent the proportions of the two subpopulations in the whole population. In this paper, we are particularly interested in scenarios where one of the subpopulations is much smaller than the other.

Definition 1 (Minority)**.** *The subpopulation with a smaller proportion in the whole population is called a minority.*

Without loss of generality, we can let the second subpopulation be the minority so that $w_1 > w_2$. In addition, we assume that the minority is contrarian; that is, their typical forecast is significantly different from the typical forecast among the majority.

Definition 2 (Contrarianism)**.** *Contrarianism implies that the distance between the typical forecasts of the two subpopulations is large relative to the standard deviations of the forecasts in each subpopulation. In other words, both* $(\mu_2-\mu_1)^2/\sigma_1^2$ *and* $(\mu_2-\mu_1)^2/\sigma_2^2$ *are large.*

Under the mixture distribution (3), the extremization transformation (2) becomes

$$a+b\mu = a+bw_1\mu_1+bw_2\mu_2. \tag{4}$$

Given that the parameters a and b cannot affect the relative weight placed on the minority, this form of extremization cannot leverage the potential of informed minorities. To address this shortcoming, we first need to identify the direction in which the minority disagrees with the population. If this could be done, we could then shift the aggregate forecast in that direction, align the aggregate more closely with the minority, and effectively allocate more weight to it.

The remainder of this subsection shows how the *skewness* of the population distribution can be used to identify the direction of the minority's disagreement. Intuitively, this is possible because a contrarian minority makes one of the tails of the population distribution heavier and hence creates a skew in that direction. One of the oldest and most popular quantifications of skewness is *Pearson's moment coefficient of skewness*:

$$\gamma = \text{skewness}(X) = \mu_3/\sigma^3 \in \mathbb{R}, \tag{5}$$

where $\sigma=\sqrt{\text{var}(X)}$ is the standard deviation and $\mu_3=\mathbb{E}\left[(X-\mu)^3\right]$ is the central third moment of the forecast population. Although the statistics literature contains alternative quantifications of skewness (such as the quantile-based statistics discussed in Groeneveld and Meeden 1984), for the remainder of this paper, we will only consider (5) and refer to it simply as *skewness*. In Section 2.3, we extend (5) to a new quantity called the *excess skewness*, in analogy to Pearson's excess kurtosis, which serves to quantify the departure of a distribution's skewness from that of a central or benchmark value.

The skewness of a mixture distribution (3) can be expressed analytically in terms of the moments of its components via a binomial expansion in the component moments. The resulting expressions, however, quickly become unwieldy as the number of components and moments increases. The expressions can be brought back under control by considering limits in which many terms in the expansion become negligibly small. The following theorem describes the skewness γ of the population mixture (3) when the minority is increasingly small and contrarian.

Theorem 1. *Consider the limit in which* w_2 *becomes small and both* $(\mu_2-\mu_1)^2/\sigma_1^2$ *and* $(\mu_2-\mu_1)^2/\sigma_2^2$ *become large. Under such an increasingly small and contrarian minority, the skewness of the mixture population of forecasts* (3) *is*

$$\gamma \approx w_2\left(\frac{\mu_2-\mu_1}{\sigma_1}\right)^3. \tag{6}$$

Proof. See Appendix A. □

Expression (6) shows that when a minority disagrees with the majority, the skewness of the population of forecasts informs us about the product of the minority's size and the direction and extent of its disagreement. In particular, if the skewness is positive (negative), then the minority believes that a positive event outcome is more (less, respectively) probable than the majority believes.

By this observation, we can now extend the extremization function with a term that adjusts the relative weight of the minority. In particular, motivated by Theorem 1, we add the cube-rooted skewness to the extremization function (4):

$$a + b\mu + c\gamma^{1/3} \approx a + bw_1\mu_1 + bw_2\mu_2 + cw_2^{1/3}\left(\frac{\mu_2 - \mu_1}{\sigma_1}\right) \tag{7}$$

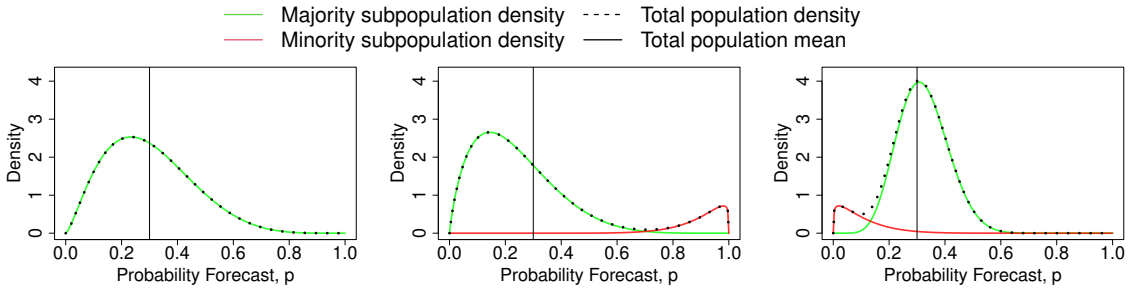
$$= a + b\left(w_1 - \frac{cw_2^{1/3}}{b\sigma_1}\right)\mu_1 + b\left(w_2 + \frac{cw_2^{1/3}}{b\sigma_1}\right)\mu_2 \tag{8}$$

for some constant c. We call this transformation the *skew-adjusted extremization*. Contrast the form (8) with that of the non-adjusted extremization in (4). By changing the value of c, the skew-adjusted extremization can either elevate ($c > 0$) or reduce ($c < 0$) the authority of the minority. Given that the values of c and the other parameters a and b are estimated from historical forecasting data, the ultimate adjustment depends on whether in the past the minority has been more or less accurate than the majority. In this way, past data tell us whether the contrarian minority should be treated as well- or ill-informed.

2.3 Probability Forecasts

Our upcoming application to real world data (see Section 3) concerns probability forecasts of future events. To describe skew-adjusted extremization in this context, we must introduce some new notation. Suppose there are n forecasters predicting the probability of a future event. The event outcome is denoted with a binary variable Y that equals 1 if the event happens; else it equals 0. Forecaster i's probability prediction $p_i \in [0,1]$ for the event $\{Y = 1\}$ is a draw from a population of forecasts, described by the density function $\pi(p)$ with mean μ, variance σ^2, and skewness γ. Before we can apply the skew-adjusted extremized-mean to these forecasts, two practical issues must be addressed.

First, Pearson's moment coefficient of skewness (5) was constructed principally to quantify the asymmetry of distributions on the whole real line. Given that in our application the forecasts represent probabilities, the population distribution of forecasts is constrained to the unit interval $[0,1]$. The positions of the interval's boundaries relative to a distribution's mean can impose an asymmetry in a way that is not related to the existence of a minority of contrarians. To approximately account for this, we perform skew-adjusted extremization based on *excess skewness* instead of skewness. The excess skewness subtracts from the skewness that which would be expected from a Beta distribution with the same mean and variance. Specifically, given that this Beta distribution



(a) $\mu = 0.3$, $\gamma_{\text{excess}} = 0$. **(b)** $\mu = 0.3$, $\gamma_{\text{excess}} = 0.75$. **(c)** $\mu = 0.3$, $\gamma_{\text{excess}} = -0.75$.

Figure 1. Population distributions with the same mean ($\mu = 0.3$) but different excess skewness. The excess skewness can quantify the size and location of the smaller mixture component relative to the larger one. Explicitly, the positive and negative values in Figures 1b and 1c reflect the minority's forecasts being higher and lower, respectively, than those of the majority.

has skewness

$$\gamma_\beta(\mu, \sigma^2) = \frac{2\sigma(1-2\mu)}{\mu(1-\mu)+\sigma^2}, \tag{9}$$

we define the excess skewness of the population distribution of forecasts $\pi(p)$ as the difference between γ and $\gamma_\beta(\mu, \sigma^2)$:

$$\gamma_{\text{excess}} := \gamma - \gamma_\beta(\mu, \sigma^2). \tag{10}$$

Figure 1 describes three different mixture populations and illustrates how the excess skewness quantifies the relative size and location of a minority of contrarian forecasters. In Figure 1a, there is no contrarian minority, the mixture aligns with a Beta distribution, and there is no excess skewness. In Figures 1b and 1c, however, there are contrarian minorities that believe that the likelihood of the event is much higher or lower, respectively, than the typical belief of the population. Given that the mixture populations in all three subplots have the same mean ($\mu = 0.3$), the mean forecast does not inform us about the presence or relative views of the minority groups. The excess skewness, however, does help us to identify such groups. In particular, it is positive when the minority's forecasts tend to be higher than those of the majority (as in Figure 1b) and negative when they tend to be lower (as in Figure 1c).

Second, directly extremizing the mean probability can exit the unit interval $[0, 1]$ and result in an aggregate forecast that cannot represent a probability. One solution is to map the probabilities to the unbounded real line with a transformation such as the logit or probit function, extremize the mean of the mapped values, and finally transform the result back to the probability scale. However, such transformations are problematic because often in practice we encounter extreme forecasts of 0 and 1 which would be mapped to $-\infty$ and $+\infty$, respectively. A typical way forward is to replace the extreme forecasts of 0 and 1 by, say, 0.01 and 0.99, respectively, before

mapping them to the real line. However, to the best of our knowledge, there is no principled way to choose the level of shrinkage. Therefore, to avoid modifying extreme forecasts, we extremize the logit of the mean probability instead of the mean of the logit-probabilities and then transform the result back to the probability scale.

Putting this all together, our final formula for aggregating probablity predictions involves a linear combination of the logit of the mean and cube root of the excess skewness of the population distribution of the probability forecasts:

$$\text{logit } \mathbb{P}(Y=1 \mid a,b,c,\mu,\gamma_{excess}) = a + b\,\text{logit}(\mu) + c\gamma_{\text{excess}}^{1/3}, \tag{11}$$

where $\text{logit}(x) := \log(x/(1-x))$ denotes the log-odds of a probablity $x \in (0,1)$. Estimating the true moments of the population distribution with the sample moments gives us

$$\text{logit } \mathbb{P}(Y=1 \mid a,b,c,\{p_i\}_{i=1}^n) \approx a + b\,\text{logit}(\hat{\mu}) + c\hat{\gamma}_{excess}^{1/3}, \tag{12}$$

where the estimated mean $\hat{\mu}$ and excess skewness $\hat{\gamma}_{excess}$ are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n p_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (p_i-\hat{\mu})^2, \qquad \hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^n (p_i-\hat{\mu})^3, \tag{13}$$

$$\hat{\gamma} = \frac{\hat{\mu}_3}{\hat{\sigma}^3}, \qquad \hat{\gamma}_{\text{excess}} = \hat{\gamma} - \frac{2\hat{\sigma}(1-2\hat{\mu})}{\hat{\mu}(1-\hat{\mu}) + \hat{\sigma}^2}. \tag{14}$$

The remaining parameters a, b, and c are estimated based on a training set of past event outcomes and their probability forecasts. Specifically, suppose we have a dataset of probability forecasts of K events whose outcomes are known. In these data, identities and numbers of forecasters of each event can be different. If $\hat{\mu}_k$ and $\hat{\gamma}_{excess,k}$ are the estimated mean and excess skewness of forecasts of the kth event outcome $Y_k \in \{0,1\}$, then the estimates $\hat{a}$, $\hat{b}$, and $\hat{c}$ can be found by fitting a logistic regression model with $\{Y_k\}_{k=1}^K$ as the response variable and the corresponding estimated moment quantities $\{\text{logit}(\hat{\mu}_k), \hat{\gamma}_{excess,k}^{1/3}\}_{k=1}^K$ as the covariates. To aggregate forecasts of a new future event k^*, we estimate the mean $\hat{\mu}_{k^*}$ and excess skewness $\hat{\gamma}_{excess,k^*}$ based on the individual forecasts and calculate the final skew-adjusted extremized-mean with $\text{logit}^{-1}(\hat{a} + \hat{b}\,\text{logit}(\hat{\mu}_{k^*}) + \hat{c}\hat{\gamma}_{excess,k^*}^{1/3})$.

3 Data Analysis

3.1 Scoring and Competing Aggregators

In this section, we test the efficacy of the skew-adjusted extremized-mean relative to a selection of its competitors. Throughout we measure accuracy with the average Brier score (Brier et al., 1950). The average Brier score of a forecaster who has predicted the chances of K events is

$$\text{BrS}\left(\{Y_k, q_k\}_{k=1}^K\right) = \frac{1}{K}\sum_{k=1}^K (Y_k - q_k)^2,$$

where $Y_k \in \{0,1\}$ is the outcome of the kth event and $q_k \in [0,1]$ is the forecaster's probability prediction of the event $Y_k = 1$. This score is negatively orientated, so that lower scores imply higher accuracy. A perfect forecaster receives a score of 0 and a constant forecast 0.5 receives a score of 0.25. Competent real-life forecasters are likely to have scores somewhere between 0 and 0.25.

The aggregation formulae we investigate are:

1. **Mean**: the equally-weighted arithmetic mean of the probability forecasts (Stone, 1961). Although this is arguably deprecated in the context of the GJP data in favour of the extremized-mean, it still provides a useful benchmark against which to judge improvements in predictive accuracy.

2. **S-trim**: the symmetrically trimmed arithmetic mean of the probability forecasts (Jose et al., 2014). This aggregator is computed as the equally-weighted arithmetic mean of the central $(1-2\kappa)\times 100\%$ of forecasts, where $\kappa \in [0,0.5]$ is the level of trimming. Preliminary analyses show that aggressive trimming with $\kappa = 0.4$ leads to the near-optimal average Brier scores for the trimmed mean. We use this value in the analyses below.

3. **HD-trim**: the highest density trimmed arithmetic mean of the probability forecast. This aggregator is computed as the equally-weighted arithmetic mean of the forecasts that lie in the shortest interval containing a certain fraction of the forecasts. For the analyses below, we specify fractions of 50% and 25%, which, according to preliminary calculations with the GJP and GK data, respectively, approximately optimize the performance of this aggregator. If S-trim is thought of as a compromise between the median and the mean, HD-trim can be thought of as a compromise between the mode and the mean. Further discussion of conceptual and computational properties of highest density regions can be found in Hyndman (1996).

4. **Votes**: the equally-weighted arithmetic mean of fully extremized probability forecasts. To compute this aggregator, we first round each individual forecast to its nearest integer (0 or 1) and then calculate the arithmetic mean of these extremized values. The appeal of this procedure, which is investigated in greater generality in Turner et al. (2014), lies in its simplicity and its relationship to voting systems. In the comparison, it provides a benchmark in our progression to more sophisticated forecast aggregation procedures.

5. **Logit**: the equally-weighted arithmetic mean of the logit-probabilities (Satopää et al., 2014). Given that the log-odds of extreme probablities 0 and 1 are infinite (recall Section 2.3), we must shrink extreme forecasts slightly towards 0.5 before calculating their logits. In this paper, we transform all forecasts linearly by shrinking the distance between each probability and 0.5 by 0.1%. The ith transformed forecast is $p_i' = 0.5 + 0.999(p_i - 0.5)$. These forecasts are then translated to log-odds and averaged with the arithmetic mean. Finally, the average log-odds is mapped back to $[0,1]$ with the inverse-logit function.

6. **Sct-crowd**: the equally-weighted arithmetic mean of forecasts from the best five participating forecasters.(Mannes et al., 2014). To calculate this select crowd aggregator, we rank the forecasters based on their average Brier scores on the training data.

7. **E-mean**: the extremized arithmetic mean of the probability forecasts (Satopää et al., 2014; Baron et al., 2014). This aggregator arises from (12) by fixing $c = 0$ and hence represents our proposed aggregator without the skew-adjustment. It adjusts both the benchmark and the level of extremization (recall Section 2.1) by estimating a and b from past forecasting data.

8. **SkE-mean**: the skew-adjusted extremized arithmetic mean of the probability forecasts. This is our proposed aggregator whose computation is described at the end of Section 2.3.

The parameters $\{a,b\}$ for E-mean and $\{a,b,c\}$ for SkE-mean are estimated from available forecasting data. However, using the same data in estimation and model evaluation is likely to result in over-fitting and underestimation of out-of-sample errors. A well-known solution is provided by cross-validation that splits the data into F approximately equal parts, known as *folds*. We separate one fold at a time, estimate the parameters from data in the other $F-1$ folds, and then, using the estimated parameters, record the accuracy of the aggregator in predicting the events in the separated fold. Cycling through all folds in this manner allows us to compute an out-of-sample error for all events in our dataset. In our analysis, we use $F = 10$ folds.

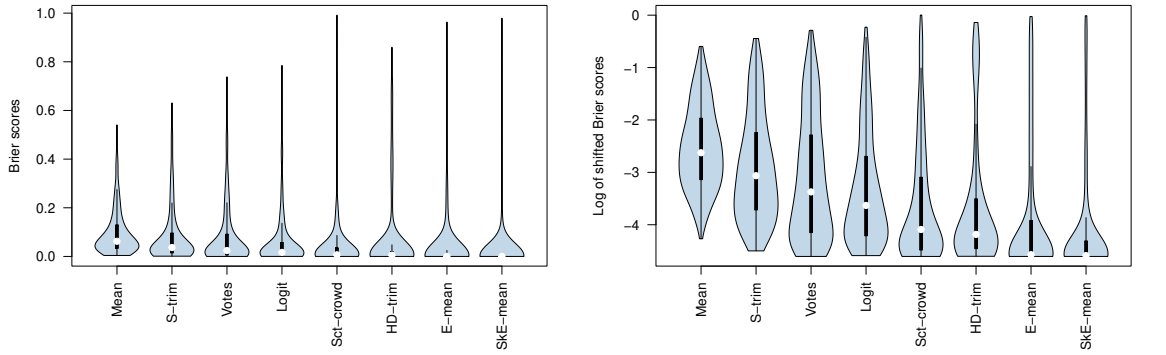
3.2 Geopolitical Events

The Good Judgment Project (GJP) data[1] include probability forecasts of 462 geopolitical events. Around 20% of these events occurred, and each event was forecast by an average of around 500 individuals. For approximately 91% of the questions, the equally-weighted arithmetic mean of the forecasts was on the correct side of 0.5. Extremizing the mean directly away from 0.5 for these cases would improve the Brier score. However, given that the Brier score penalizes larger errors more heavily, extremizing the mean all the way to the nearest integer (0 or 1) is not the overall most accurate strategy. Instead, in line with Baron et al. (2014), we observe that E-mean with an extremizing parameter of $b \approx 3$ leads to better results and improves the average Brier score of the (un-extremized) mean aggregate by around 40%.

Figure 2 presents violin plots of the Brier scores across all questions for each aggregation procedure. In particular, Figure 2a describes the raw Brier scores. To better visualize the differences, Figure 2b plots the log of (slightly) shifted Brier scores. In each plot, the white dots represent the median scores and the black rectangles inside the violin plots describe the interquartile ranges. The widths of the blue shapes are proportional to smoothed empirical densities of the scores.

The results display an inverse relationship between mean and variance of the individual scores. In particular, the more accurate aggregators tend to make more extreme errors. This illustrates an inherent trade-off in

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



(a) Raw Brier scores

(b) Brier scores transformed with the function $\log(x+0.01)$

Figure 2. Violin plots of the Brier scores of different aggregators on the GJP forecasting data. The widths of the blue shapes are proportional to smoothed empirical densities of the scores. The white dots mark the median scores (for the means scores, see Table 2 in Appendix B) and black rectangles mark interquartile ranges.

probabilistic forecasting that often seeks to maximize sharpness (i.e., extremity) subject to calibration (i.e., compatibility with the empirical frequency of the outcomes; Ranjan and Gneiting 2010). Such a strategy maximizes performance in terms of the Brier score but at the risk of occasionally making large errors. For instance, compare the following two calibrated forecasts: a naive forecast that is always equal to 0.5, and a well-informed forecast that alternates in equal proportion between 0.1 and 0.9. The error of the naive forecast is always 0.25. The error of the well-informed forecast is small 0.01 with 90% chance but large 0.81 with 10% chance. Even though the well-informed forecast occasionally makes large errors, it is generally considered more useful than the naive forecast because it can discriminate among the different outcomes of the event.

Among all competing aggregators, E-mean and SkE-mean perform the best. Even though their median scores look almost identical, the interquartile ranges of their Brier scores reveal that SkE-mean achieves very low Brier scores more often than E-mean. In Appendix B we consider the mean Brier scores and show that the mean Brier score of SkE-mean (0.046) is approximately 25% smaller than the mean Brier score of E-mean (0.061). This difference is smaller than but of the same order of magnitude as the reduction from the mean Brier score of Mean (0.098) to that of E-mean. The Brier scores of E-mean and SkE-mean are smaller than those of Mean for approximately 85% and 90% of the GJP events, respectively.

Table 1a presents the estimates of the model parameters used in E-mean and SkE-mean. The extremizing parameter in E-mean, namely the slope coefficient on $\text{logit}(\hat{\mu})$, is approximately 3.4. However, if we include the skew-adjustment and consider SkE-mean, this slope coefficient increases to 4.8. We can then extremize the logit of the mean probability more aggressively if the skewness of the population is accounted for. As intuition would suggest, when a well informed minority agrees with the majority, we ought to be more confident in that

(a) Good Judgment Project Data

	Estimated Coefficients	
Covariate	E-mean	SkE-Mean
Intercept	-0.415^{**}	-0.162
	(0.195)	(0.233)
$\text{logit}(\hat{\mu})$	3.442^{***}	4.878^{***}
	(0.371)	(0.525)
$\hat{\gamma}^{1/3}_{excess}$		3.080^{***}
		(0.583)
Observations	462	462
Log Likelihood	-91.765	-71.352
Akaike Inf. Crit.	187.529	148.704

(b) General Knowledge Data

	Estimated Coefficients	
Covariate	E-mean	SkE-Mean
Intercept	-2.331^{***}	-1.780^{***}
	(0.253)	(0.282)
$\text{logit}(\hat{\mu})$	4.432^{***}	4.376^{***}
	(0.401)	(0.406)
$\hat{\gamma}^{1/3}_{excess}$		1.706^{***}
		(0.377)
Observations	500	500
Log Likelihood	-163.548	-151.478
Akaike Inf. Crit.	331.096	308.956

Note: *p<0.1; **p<0.05; ***p<0.01

Table 1. Summary statistics of the fitted logistic regression models under the GJP and GK data. The values in the parentheses are standard errors. The stars describe the p-values of hypothesis tests of the parameters being exactly zero.

majority. Seen from another perspective, we can afford to be bolder with mean-extremization given that we have the skew-adjustment to temper its effects when the majority view is challenged by an informed minority.

Table 1a also shows the estimated slope coefficient of $\hat{\gamma}^{1/3}_{excess}$ in SkE-mean. Depending on whether $\hat{\gamma}^{1/3}_{excess}$ is positive or negative, SkE-mean treats the minority as well- or ill-informed, respectively (recall end of Section 2.2). Given that the estimate of this slope coefficient is positive and statistically significant, SkE-mean expects the minority in the GJP data to be well-informed. To illustrate the effect of this, Figure 3 presents forecast distributions for the 20 forecasting questions for which E-mean and SkE-mean differ the most. The questions themselves are listed in Table 6 in Appendix C. For many of these distributions E-mean is far from the realized outcome while SkE-mean is often, although not consistently, closer. Furthermore, many of the presented distributions are characterized by a spike at the side of the observed outcome. The spike can be taken to represent a contrarian minority that co-exists alongside a more diffuse majority. Given that SkE-mean treats the minority as well-informed, it allocates more importance to the minority. This often places it between E-mean and the consensus belief of the minority (consider, e.g., questions 1, 6, 7, or 8 in Figure 3). To the extent that the minority continues to be informed in future forecasting questions, this strategy will pay off and allow SkE-mean to outperform E-mean.

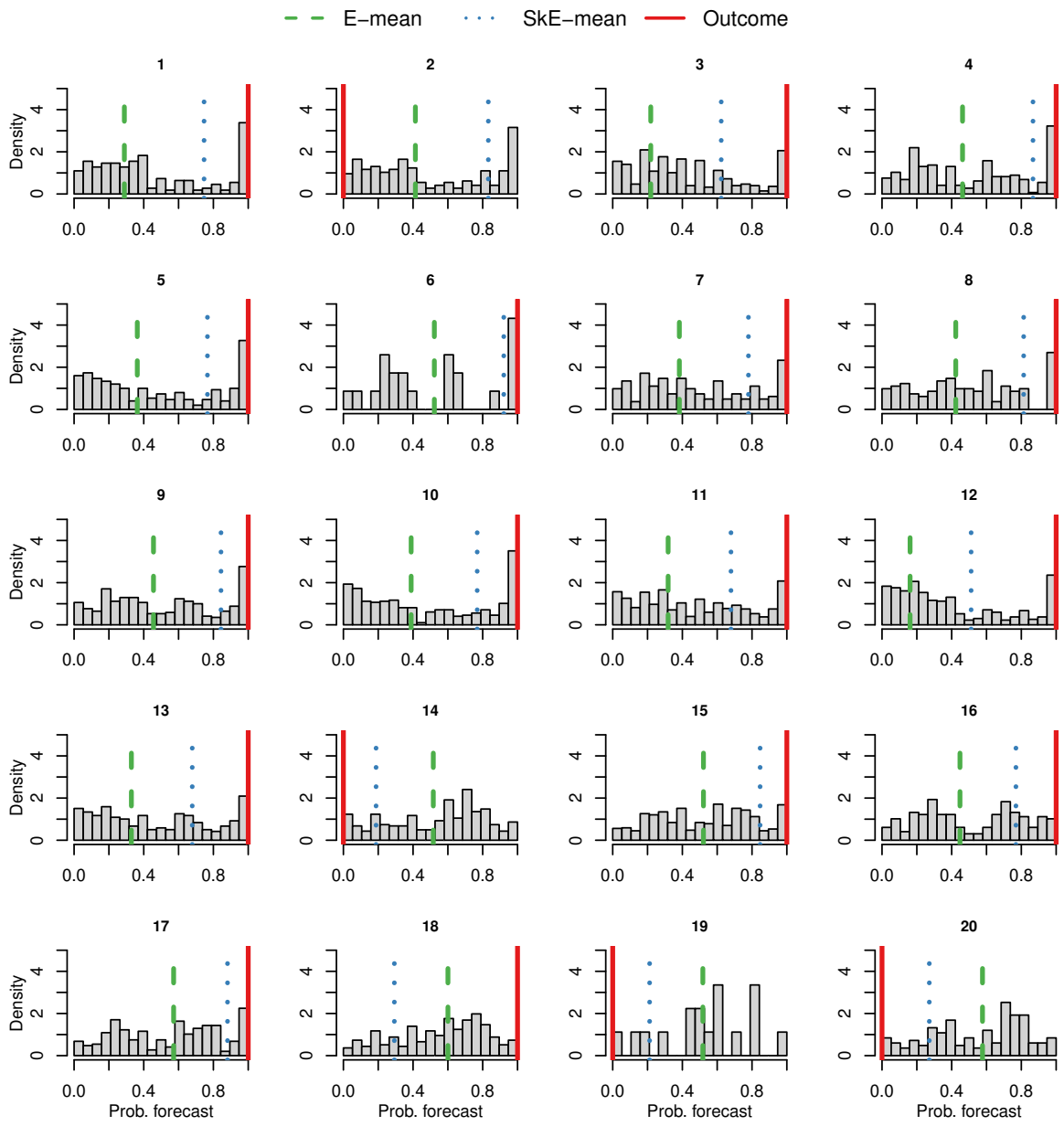
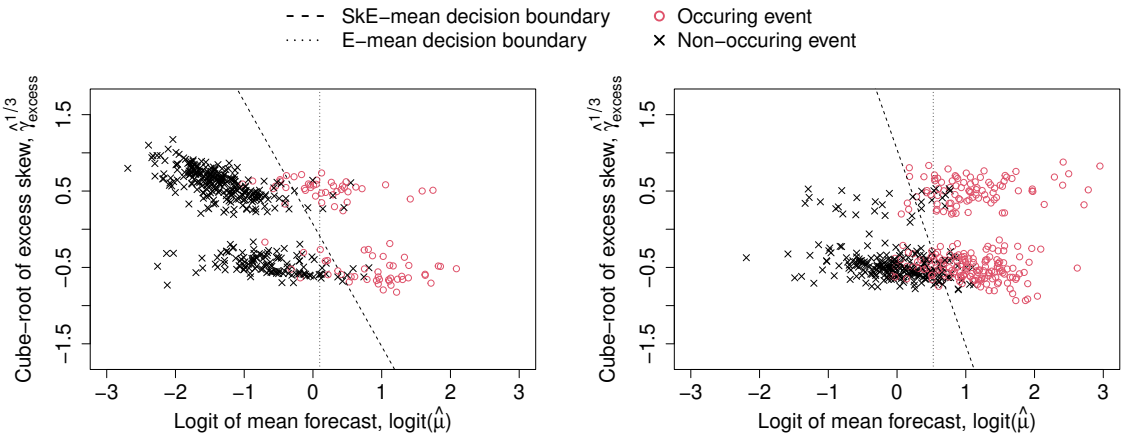


Figure 3. Forecast of the 20 GJP events for which E-mean and SkE-mean differ the most. In most cases SkE-mean is closer to the observed outcome and between a contrarian minority and E-mean.

By placing relatively more weight on the minority, skew-adjustment is not merely adjusting the level of extremization. Instead, it can reverse the direction of extremization and point towards different outcomes altogether. To illustrate, Figure 4a describes all questions in the GJP dataset by circles whose coordinates are given by the logit of the mean $\text{logit}(\hat{\mu})$ and the cube-root excess skewness $\hat{\gamma}_{excess}^{1/3}$ of the forecasts. Assuming that type 1 and 2 errors are equally costly, the vertical and downward sloping lines show the decision boundaries



(a) Good Judgment Project **(b)** General Knowledge

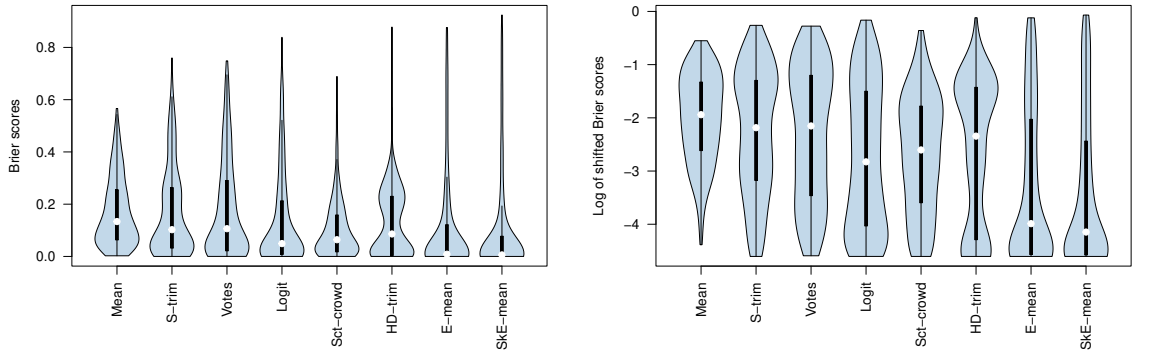
Figure 4. Potential decision boundaries for the E-mean and SkE-mean. The lines show where the aggregators predict a 50% chance of event occurrence. Any point on the right (left) hand side of a decision boundary would then result in an aggregated forecast greater (lower, respectively) than 0.5. The red points to the right of the downward sloping line but to the left of the vertical line represent the questions for which SkE-mean points to the right answer when E-mean does not. Counts of points in each region are provided in tables 4 and 5.

for E-mean and SkE-mean, respectively. Specifically, the lines show the values of logit($\hat{\mu}$) and $\hat{\gamma}^{1/3}_{excess}$ for which the formulae would produce an aggregated forecast of 0.5. Any point on the right (left) hand side of a decision boundary would then result in an aggregated forecast greater (lower, respectively) than 0.5. The circles in the two wedges formed by the two decision boundaries represent questions for which E-mean and SkE-mean fall on different sides of 0.5 and hence point towards different outcomes. Looking at Figure 4a, the decision boundary of SkE-mean seems to do a better job at separating events that occurred from those that did not. In fact, E-mean and SkE-mean are on the 'right' side of 0.5 in 91% and 95% of the cases, respectively (see Tables 4–5 in Appendix B for the empirical results informing these values). Therefore, SkE-mean points towards the correct outcome more often and hence has the potential to improve decision making. We note also that the apparent clustering of points in Figure 4 arises from a tendency of the inferred minority to take an extreme position one way or another. The inter-question differences in the resulting excess skews is then amplified by the cube-root function that, very approximately, acts like a step function centred at zero pushing apart the negative and positive values.

3.3 General Knowledge Questions

Martinie et al. (2020) consider 500 general knowledge statements and collect approximately 100 probabilistic forecasts of each statement being true.[2] Half of the statements are true and for approximately 75% of them the mean probability forecast is on the correct side of 0.5. As to before, partial extremizing yields better accuracy

[2]The data can be downloaded at https://doi.org/10.1371/journal.pone.0232058.s003



(a) Raw Brier scores

(b) Brier scores transformed with the function $\log(x+0.01)$

Figure 5. Violin plots of the Brier scores of different aggregators on the GK forecasting data. The widths of the blue shapes are proportional to smoothed empirical densities of the scores. The white dots mark the median scores (for the means scores, see Table 3 in Appendix B) and black rectangles mark interquartile ranges.

than full extremization. Specifically, without considering skew-adjustment, the optimal level of extremizing is given by $b \approx 4$.

Figure 5 presents the violin plots of the average Brier scores for each aggregation procedure. In particular, Figure 5a plots the raw Brier scores and Figure 5b plots the log of (slightly) shifted Brier scores. Among all competing aggregators, SkE-mean performs best and E-mean performs second best in terms of median Brier score. In Appendix B, we present the mean Brier scores and show that the mean Brier score of SkE-mean (0.098) is approximately 10% smaller than the mean Brier score of E-mean (0.107). This reduction is smaller than that under the GJP data both in relative and absolute terms. The Brier scores of E-mean and SkE-mean are smaller than those of Mean for approximately 76% and 79% of the GK questions, respectively.

Table 1b describes the estimates of the model parameters used in E-mean and SkE-mean. The estimates are similar to those under the GJP data (recall Table 1a). In fact, only the intercept terms are significantly different. Given that the intercept under the GK data is moderately large negative, the GK forecasts are systematically too high. In fact, close inspection of the data reveals that the mean forecast is greater than 0.5 for approximately 75% of the questions. Given that only 50% of the statements are true, the forecasters exhibit a noticeable upward bias. A potential explanation is that the general knowledge statements were constructed by taking true facts and modifying, or negating, key words in a subset of them. As a consequence, the statements use credible, impressive-sounding terminology that may act as an informal proxy for truth among the forecasters. Several statements, however, have been subtly modified to make them false in a way that only an expert would identify. The reverse trick – constructing a silly or dubious statement which is nonetheless true – is harder to achieve when constructing questions. This may be why we see a greater proportion of forecasts at moderately high values. The

negative intercept in both E-mean and SkE-mean tries to correct for this upward bias.

The coefficient on the cube-root excess skew $\hat{\gamma}_{excess}^{1/3}$ is again positive and statistically significant, suggesting that the minority be deemed well-informed and an important contributor to the accuracy of the final aggregate forecast. Even though this estimate is smaller than that under the GJP data, SkE-mean is still picking up on idiosyncratic characteristics that we associate with the contrarian minorities. To illustrate, Figure 6 presents forecast distributions for the 20 forecasting questions for which E-mean and SkE-mean differ the most. The tendency towards a contrarian minority is more visible in questions 3, 5, and 7.

Martinie et al. (2020) rated their questions from level 1 to 5 in terms of difficulty. The easiest, level-1 questions include statements such as 'The moon shines at night because it reflects light from the sun,' whereas the hardest, level-5 questions include statements such as 'Microwaves contain more energy than visible light'. The questions for which E-mean and SkE-mean differ the most tend to be very difficult: among the 20 questions with the largest disagreement between these two aggregators, the average difficulty rating is 4.4. Indeed, a reasonable definition of a 'difficult question' in this context could require the presence of a poorly-informed majority and a well-informed contrarian minority.

Finally, the vertical and downward sloping lines in Figure 4b present the decision boundaries for E-mean and SkE-mean, respectively. Again, the underlying assumption is that type 1 and 2 errors are equally costly so that 0.5 acts as a threshold between 'action' and 'no action.' Even though the decision boundary of SkE-mean is better at separating the true outcomes, the difference is not as large as it was under the GJP data: this time E-mean and SkE-mean are on the 'right' side of 0.5 in 85% and 86% of the cases, respectively.

4 Discussion

4.1 Theoretical Suggestions

In this paper, we analyzed two data sets of contrasting forecasting problems and showed that the excess skewness of a distribution of forecasts is a significant positive predictor of event occurrence. The resulting skew-adjusted extremized-mean, i.e, SkE-mean outperforms its competitors, including the arithmetic mean, trimmed mean, mean of the five most accurate forecasters, and several other commonly used aggregators. To reinforce generalizability of our findings, we presented a theoretical argument for the excess skewness acting as a proxy for the presence of a question-specific minority of contrarian forecasters. Past outcomes and their forecasts are then used as training data to determine whether the minority should be given more or less relative weight in future aggregation problems. In both of our empirical studies, the minority was found to be informed, which aligns with the usual assumptions made in the 'contrarian minority' literature (Prelec et al., 2017; Martinie et al., 2020; Palley and Satopää, 2021). The empirical findings largely speak for themselves, and we encourage readers to perform their own analyses of the publicly available data to reconfirm them.

The skew-adjustment appears to be advantageous for both geopolitical and general knowledge forecasting

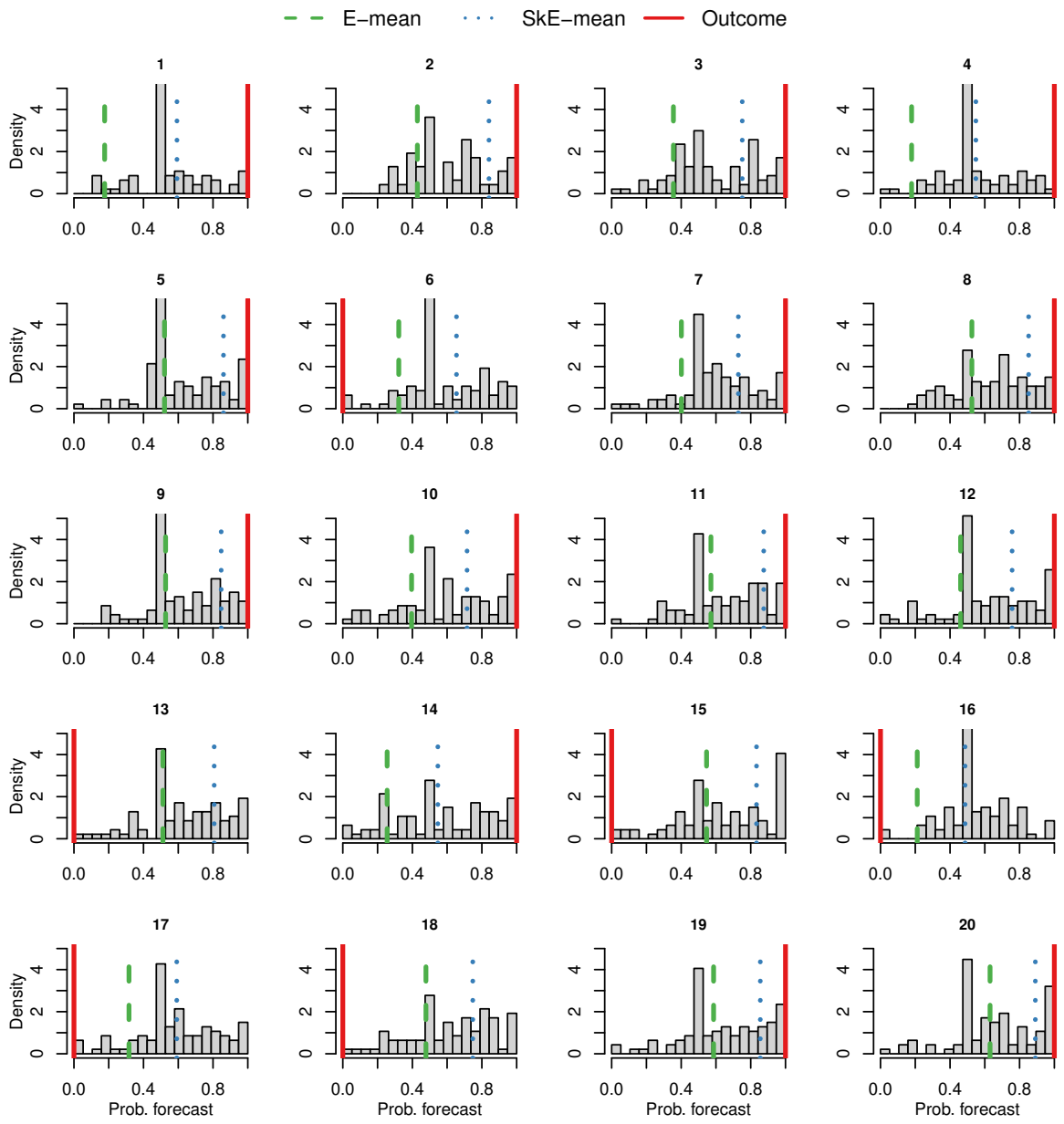


Figure 6. Forecast for the 20 GK questions for which E-mean and SkE-mean differ the most. In most cases SkE-mean is closer to the observed outcome.

problems, albeit to different degrees. Furthermore, the optimal level of adjustment appears to be similar in the two cases, as can be observed from the similarity between the fitted logistic regression coefficients in Table 1 and the resulting decision boundaries in Figure 4. Whether or not this ought to be considered surprising is open to debate. On one hand, both contexts ask the forecasters to quantity their beliefs in terms of probability predictions. On the other hand, the latent mechanisms by which forecasters access and process data are likely to be different

in the two contexts. We conjecture that in both geopolitical and general knowledge contexts the information needed to make accurate forecasts is 'lumpy' in nature. There are, for example, discrete conclusive pieces of evidence or revelatory concepts that are hard to access and hence are only held by a minority of forecasters. The power of these pieces of information is enough to push forecasts deep into one of the tails of the forecast distribution, increasing the skew in that direction. Support for the existence and importance of such 'threshold concepts' or 'eureka moments' can be found in the psychology and educational literatures (e.g., Jones 2003 and Land et al. 2008). Their relevance to forecast aggregation, however, is less developed. Similar levels of skew-adjustment being appropriate for both contexts suggests that the distributions of knowledge or information among the forecasters share similar properties, and any work to confirm or refute this empirically would be valuable. Of particular interest would be to investigate data with uninformed or misinformed minority groups. Such minorities could be inferred from negative estimates of the coefficient c in (12).

Our SkE-mean is premised on the relevance of a non-trivial structural feature of the population of forecasters. Specifically, it relies on the existence of distinct subpopulations. One may ask why we have chosen to quantify this feature indirectly via skewness rather than employing more sophisticated statistical methodology to infer the parameters of the mixture distribution explicitly. Why, for example, did we not estimate the weights, means and variances in (3) using numerical Markov chain Monte Carlo (MCMC) methods developed for Bayesian inference? Our principal reason is computational tractability. The excess skewness of a sample is quick and easy to compute, whereas MCMC algorithms for mixture distributions are liable to become protracted as they jump (or, more problematically, do not jump) around the parameter space. Furthermore, such MCMC algorithms can be highly dependent on the likelihood functions for the latent parameters, which in our case would entail choosing precise specifications of the component distributions π_s in (3). This is a specification that we anticipate most users in practice are reluctant to make.

Our analysis only captures associations in the data. Therefore, it is possible that the empirically observed relationship between the excess skewness and event outcomes is not causally attributable to the presence of an informed minority. Given that the relationship can be verified and quantified with available data while the underlying cause, at least for now, cannot, we consider the relationship as the main focus of our work. Overall, we see the development of models for forecaster behaviour as one of the key drivers of predictive methods but not necessarily as the predictive methods themselves. For this reason, we are interested in experimenting with models of forecaster interdependence, such as the information diversity model of Satopää et al. (2016), the model of private and shared information of Lichtendahl Jr et al. (2020), and the Bayesian group belief model of Dietrich (2010). Such models can offer further insight and point towards other important summary statistics of observed forecaster behaviour.

4.2 Practical Suggestions

In terms of the log-odds, we recommend the aggregated forecast

$$\text{logit}\,\mathbb{P}(Y=1 \mid a,b,c,\{p_i\}_{i=1}^n) = a + b\,\text{logit}\,(\hat{\mu}) + c\hat{\gamma}_{\text{excess}}^{1/3}. \tag{15}$$

The coefficients in (15) have simple interpretations: a determines the reference value for extremization (see Section 2.1) and corrects for populations of forecasters whose average member consistently produces forecasts that are too high or too low; b controls the strength of extremization, so that $b > 1$ extremizes the mean away from the reference value, $b = 1$ results in the unadjusted mean, and $0 < b < 1$ shrinks the mean towards the reference value; c controls the skew-adjustment, which we have argued may be a way to account for informed minorities, so that $c > 0$ effectively up-weights the minority, $c = 0$ leaves it unweighted, and $c < 0$ down-weights it. In particular, $a = 0, b = 1, c = 0$ reverts to the unadjusted mean probability. Even though coefficient values of $a \approx -1$, $b \approx 4$ and $c \approx 2$ are found to produce low Brier scores for the GJP and GK data, we strongly encourage users of SkE-mean to estimate context-specific coefficient values from historical forecast data wherever possible.

4.3 Future Research

One future research direction involves extending SkE-mean to a multivariate setting, where forecasters predict events with $M > 2$ possible outcomes. If each forecaster i's prediction $\mathbf{p}_i$ is in the $(M-1)$-simplex and the event results in one of M outcomes $Y \in \{1,\ldots,M\}$, then one approach is to compute the mean, $\hat{\mu}_m$, and excess skewness, $\hat{\gamma}_{\text{excess},m}$ separately for the probabilities assigned to each potential outcome $m \in \{1,2,\ldots,M\}$, and then calibrate their influence on the aggregated probability using a multinomial logistic regression model. For each potential outcome $m = 1,\ldots,M$, the aggregated forecast then is

$$\mathbb{P}(Y=m \mid a,b,c,\{\mathbf{p}_i\}_{i=1}^n) = \frac{e^{x_m}}{1+\sum_{j=1}^{M-1} e^{x_j}}, \tag{16}$$

where $x_j = a + b\,\text{logit}(\hat{\mu}_j) + c\hat{\gamma}_{\text{excess},j}^{1/3}$. This approach is appealing because it capitalizes on an assumed symmetry between the M potential outcomes and hence keeps the number of parameters to be estimated (i.e., a, b, and c in Eq. 16) low. It also preserves our rationale relating to contrarian minorities because we have, in a sense, reduced the multivariate problem to $K-1$ univariate problems with a form that we have already seen.

Our goal in the present work was to show how E-mean could be usefully extended with the help of standard statistical modeling techniques, such as the logistic regression. SkE-mean, however, involves more parameters than E-mean. Given that accurate estimation of more complex models often requires more data, SkE-mean can be expected to outperform E-mean only if the training dataset is sufficiently large. Both data sets in the current paper feature large numbers of questions (with which to estimate extremization and skew adjustment parameters) and large numbers of forecasters per question (with which to estimate the excess skewness of the forecasts). As these numbers become small, estimation errors may adversely affect our proposed aggregation method. In Appendix E, we subsample both the GJP and GK data and consider the relative performance between E-mean

and SkE-mean under different sample sizes. As a simple rule of thumb, our skew-adjustment is likely to improve E-mean as long as there are at least 30 events and 30 forecasts per event.

Even though we do not consider these data requirements unreasonable, a future research project could look for ways to improve the performance of SkE-mean under small data. A solution is likely to employ procedures that shrink parameter estimates towards default values. Particularly relevant techniques are kernel density estimation and regularized logistic regression, which are described in Sections 6.6.1 and 4.4.4 of Hastie et al. (2001), respectively.

References

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

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

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

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

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

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.

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*, 2(4):1360–1383.

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

Groeneveld, R. A. and Meeden, G. (1984). Measuring skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 33(4):391–399.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126.

Jones, G. (2003). Testing two cognitive theories of insight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5):1017.

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

Karmarkar, U. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance*, 21(1):61–72.

Land, R., Meyer, J. H., and Smith, J. (2008). *Threshold concepts within the disciplines*. Sense Publishers.

Laplace, P. S. (1774). Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*.

Lattimore, P. K., Baker, J. R., and Witte, A. D. (1992). The influence of probability on risky choice: A parametric examination. *Journal of economic behavior & organization*, 17(3):377–400.

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., Jose, V. R. R., and Winkler, R. L. (2020). Extremizing and anti-extremizing in bayesian ensembles of binary-event forecasts. Available at SSRN: https://ssrn.com/abstract=2940740.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). *Calibration of Probabilities: The State of the Art*, pages 275–324. Springer Netherlands, Dordrecht.

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276.

Martinie, M., Wilkening, T., and Howe, P. D. (2020). Using meta-predictions to identify experts in the crowd when past performance is unknown. *Plos one*, 15(4).

Palley, A. and Satopää, V. A. (2021). Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. *Available at SSRN: https://ssrn.com/abstract=3504286*.

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

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

Ravishanker, N., Chi, Z., and Dey, D. K. (2002). *A first course in linear model theory*. Chapman and Hall/CRC.

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

Satopää, V. A., Jensen, S. T., Pemantle, R., and Ungar, L. H. (2017). Partial information framework: Model-based aggregation of estimates from diverse information sources. *Electronic Journal of Statistics*, 11(2):3781–3814.

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

Satopää, V. A. (2021). Improving the wisdom of crowds with analysis of variance of predictions of related outcomes. *International Journal of Forecasting*, 37(4):1728–1747.

Sheynin, O. B. (1977). Laplace's theory of errors. *Archive for history of exact sciences*, 17(1):1–61.

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

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

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*.

Zellner, M., Abbas, A. E., Budescu, D. V., and Galstyan, A. (2021). A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science*, 8(2):201187.

A Appendix: A Limiting Form for the Skewness

This appendix derives the approximate expression (6). First, the jth moment of a two-component mixture distribution can be written as a double-sum, where the external sum marginalizes over the component distributions and the internal sum expands differences from the distribution mean in terms of differences from the component means. More precisely,

$$\begin{aligned}
\mathbb{E}[(X-\mu)^j] &= \sum_{s=1}^{2} w_s \, \mathbb{E}[(X-\mu)^j \mid X \sim \pi_s] \\
&= \sum_{s=1}^{2} w_s \, \mathbb{E}[(X-\mu_s+\mu_s-\mu)^j \mid X \sim \pi_s] \\
&= \sum_{s=1}^{2} w_s \sum_{k=0}^{j} \binom{j}{k} (\mu_s-\mu)^{j-k} \, \mathbb{E}[(X-\mu_s)^k \mid X \sim \pi_s].
\end{aligned}$$

Given that the differences between the component means and the population mean can be written as $\mu_1-\mu = \mu_1-(w_1\mu_1+w_2\mu_2) = w_2(\mu_1-\mu_2)$ and $\mu_2-\mu = \mu_2-(w_1\mu_1+w_2\mu_2) = (1-w_2)(\mu_2-\mu_1)$, we can express the second and third central moments of the mixture distribution in terms of properties of the mixture components

only. Furthermore, we can investigate how these quantities behave as w_2 becomes small. Denoting the first, second, and third central moments of the subpopulation s by μ_s, σ_s^2 and $\mu_{3,s}$ respectively, the third central moment of the mixture distribution is given by

$$\begin{aligned}
\mathbb{E}[(X-\mu)^3] &= \sum_{s=1}^{2} w_s \left[(\mu_s-\mu)^3 + 3(\mu_s-\mu)^2\,\mathbb{E}[(X_s-\mu_s)] + 3(\mu_s-\mu)\,\mathbb{E}[(X_s-\mu_s)^2] + \mathbb{E}[(X_s-\mu_s)^3]\right] \\
&= \sum_{s=1}^{2} w_s \left[(\mu_s-\mu)^3 + 3(\mu_s-\mu)\sigma_s^2 + \mu_{3,s}\right] \\
&= (1-w_2)w_2^3(\mu_1-\mu_2)^3 + 3(1-w_2)w_2(\mu_1-\mu_2)\sigma_1^2 + (1-w_2)\mu_{3,1} \\
&\quad + w_2(1-w_2)^3(\mu_2-\mu_1)^3 + 3w_2(1-w_2)(\mu_2-\mu_1)\sigma_2^2 + w_2\mu_{3,2} \\
&= w_2(\mu_2-\mu_1)^3 + O(\sigma_1^2) + O(\sigma_2^2) + O(w_2^2) \quad \text{as} \quad \sigma_1^2 \to 0,\ \sigma_2^2 \to 0,\ w_2 \to 0, \qquad (17)
\end{aligned}$$

where the products involving third central moments are subsumed into the $O(\sigma_s^2)$ terms. This follows from the expectations of the cubes of random variables in $[0,1]$ being smaller than the expectations of their squares. Similarly considering the second central moment of the mixture, we have

$$\mathbb{E}[(X-\mu)^2] = \sigma_1^2 + O(w_2) \quad \text{as} \quad w_2 \to 0. \qquad (18)$$

Under the assumption that $|\mu_2-\mu_1| > 0$ and that w_2 decreases faster than σ_1^2, equations (17) and (18) give us the final limiting expression:

$$\lim_{\sigma_1^2\to 0,\sigma_2^2\to 0, w_2/\sigma_1^2\to 0} \frac{\mathbb{E}[(X-\mu)^3]}{(\mathbb{E}[(X-\mu)^2])^{3/2}} = w_2(\mu_2-\mu_1)^3/\sigma_1^3.$$

B Appendix: Tabulated Results

Aggregator	Mean	S-trim	Votes	HD-trim	Logit	E-mean	Sct-crowd	SkE-mean
Average	0.098	0.084	0.080	0.073	0.063	0.061	0.054	0.046
Standard Deviation	0.095	0.118	0.122	0.171	0.117	0.169	0.133	0.159

Table 2. Summary statistics for the out-of-sample Brier scores achieved by different aggregators on the Good Judgment Project data. For descriptions of the aggregators, see Section 3.1.

Aggregator	Mean	S-trim	Votes	HD-trim	Logit	E-mean	Sct-crowd	SkE-mean
Average	0.171	0.171	0.179	0.131	0.135	0.107	0.107	0.098
Standard Deviation	0.131	0.170	0.188	0.133	0.177	0.197	0.120	0.197

Table 3. Summary statistics for the out-of-sample Brier scores achieved by different aggregators on the general knowledge forecasting data. For descriptions of the aggregators, see Section 3.1.

	E-mean right	E-mean wrong
SkE-mean right	414	20
SkE-mean wrong	8	20

Table 4. Counts of Good Judgment Project questions for which E-mean and SkE-mean are on the right or wrong side of 0.5.

	E-mean right	E-mean wrong
SkE-mean right	415	17
SkE-mean wrong	11	57

Table 5. Counts of general knowledge questions for which E-mean and SkE-mean are on the right or wrong side of 0.5.

C Appendix: Good Judgment Project Questions

1	Will Fayez al-Tarawneh resign or otherwise vacate the office of Prime Minister of Jordan before 1 January 2013?
2	Will the sentence of any of the three members of the band Pussy Riot who were convicted of hooliganism be reduced, nullified, or suspended before 1 December 2012?
3	Will Viktor Yanukovich vacate the office of President of Ukraine before 10 May 2014?
4	Will Christian Wulff resign or vacate the office of President of Germany before 1 April 2012?
5	Will Viktor Yanukovich vacate the office of President of Ukraine before 10 May 2014?
6	Will China officially announce a peak year for its carbon emissions before 1 June 2015?
7	Before 1 April 2014, will one or more countries impose a new requirement on travelers to show proof of a polio vaccination before entering the country?
8	Will the official US Dollar to Venezuelan Bolivar exchange rate exceed 4.35 at any point before 1 April 2013?
9	Will the Malian government and Ansar Dine commence official talks before 1 April 2013?
10	Will a foreign or multinational military force fire on, invade, or enter Syria between 6 March 2012 and 31 December 2012?
11	Before 1 February 2014, will Iran officially announce that it has agreed to significantly limit its uranium enrichment process?
12	Will the official US Dollar to Venezuelan Bolivar exchange rate exceed 4.35 at any point before 1 April 2013?
13	Will Goodluck Jonathan vacate the office of President of Nigeria before 10 June 2015?
14	Will Israel officially establish a date for early elections before 6 November 2012?
15	Will China conduct naval exercises in the Pacific Ocean beyond the first island chain before 1 June 2015?
16	Will the IMF officially announce before 1 January 2013 that an agreement has been reached to lend Egypt at least 4 billion USD?
17	Will the IMF officially announce before 1 January 2013 that an agreement has been reached to lend Egypt at least 4 billion USD?
18	Will Serbia be officially granted EU candidacy before 1 April 2012?
19	By 31 December 2011, will the World Trade Organization General Council or Ministerial Conference approve the 'accession package' for WTO membership for Russia?
20	Will the Israeli-Palestinian peace talks be extended beyond 29 April 2014?

Table 6. Questions used to elicit the forecasts shown in Figure 3

D Appendix: General Knowledge Statements

1	There are four covalent bonds involved in a methane molecule.
2	An increase in current through a wire exposed to a magnetic field will also increase the force experienced by the wire.
3	Photosynthesis is an example of an endothermic reaction.
4	The first two electron shells in Neon are fully filled with electrons.
5	In physics, U-values measure how effective a material is an insulator.
6	Nitrogen can typically form up to two covalent bonds.
7	Knowing an appliance's power consumption and potential difference would allow someone to calculate the current.
8	Hedgehogs are nocturnal and hibernate during the winter.
9	In an ammonia molecule, hydrogen and nitrogen atoms share electrons.
10	As a substance changes state from liquid to gas, the amount of energy particles have increases.
11	The force experienced by a current-carrying wire can be reversed by reversing the direction of current/magnetic field.
12	Isotopes have the same number of protons, but different number of neutrons.
13	Eye color is an example of continuous variation in a trait.
14	The last ice age occurred during the Jurassic period.
15	In a circuit, a fuse can be reset after it is triggered.
16	Sound waves and electromagnetic waves are examples of longitudinal waves.
17	Secondary industries dominate the market in emerging economies.
18	As the temperature increases, the solubility of gasses increases.
19	In physics, work done is equal to the force needed to move an object multiplied by the distance it moved.
20	If the voltage in a circuit remains constant but the resistance is increased, current decreases.

Table 7. General knowledge statements that were used to elicit the forecasts shown in Figure 6

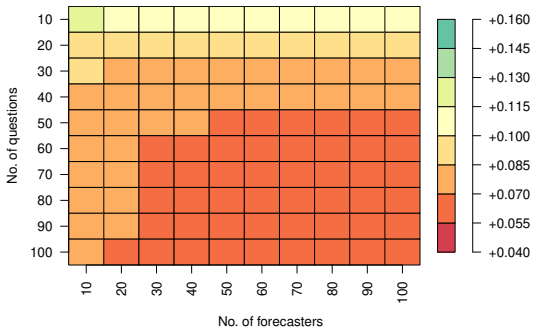
E Appendix: Sample Size Requirements

Given that accurate estimation of more complex models often requires more data, SkE-mean can be expected to outperform E-mean only if the training dataset is sufficiently large. To understand the data requirements of our skew adjustment, we sub-sample the GJP and GK data and compare the performance of E-mean and SkE-mean given a smaller number of events/questions and forecasts per event.

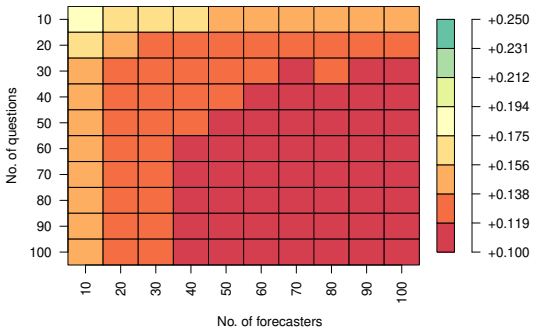
Figure 7 presents the results in a series of levelplots. In each levelplot, the x-axis represents the number of forecasters per event and the y-axis represents the number of events/questions in the training dataset. The left column considers GJP data, and the right column considers GK data. Figures 7a–7b in the top row present the average (out-of-sample) Brier scores of E-mean. The corresponding plots for SkE-mean are given by Figures 7c and 7d in the middle row. Each value is an average (out-of-sample) score over 500 random partitions of the data into training data (with which to estimate the model parameters) and testing data (with which to compute out-of-sample Brier scores). Figures 7e–7f in the bottom row show the average Brier score of SkE-mean divided by the average Brier score of E-mean. Therefore, a value smaller (larger) than 1.0 indicates that SkE-mean performs, on average, better (worse, respectively) than E-mean.

According to the results, both procedures perform poorly under a small number questions (e.g., 10 questions). This happens because if we randomly draw a small number of questions from the full data, the training data can end up being perfectly separated in the sense that the logit of the average probability always falls on the right side of some reference value μ_0. Under such perfect separation, the most 'extreme extremizing' is the most appropriate (i.e., the optimal value of the extremization constant is $\alpha = +\infty$). However, if the testing data include events for which the average forecast is on the wrong side of the reference value, 'extreme extremizing' leads to correspondingly extremely wrong predictions and hence to poor out-of-sample accuracy. To alleviate this problem, it may be possible to regularize extremization with an appropriate constraint or Bayesian prior (Gelman et al., 2008).

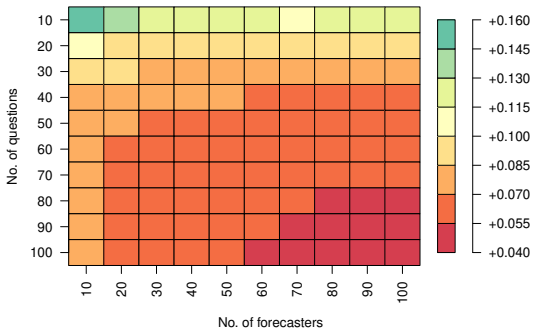
Even though both E-mean and SkE-mean face the dangers of over-fitting under small data, SkE-mean has one extra degree of freedom and hence can be more susceptible to overfitting. As a result, E-mean outperforms SkE-mean when both the number of questions and forecasts per event is small. However, as either dimension of the data increases, SkE-mean outperforms E-mean by an increasing margin. In particular, SkE-mean improves upon E-mean when the training data include at least 30 events and 30 forecasts per event.



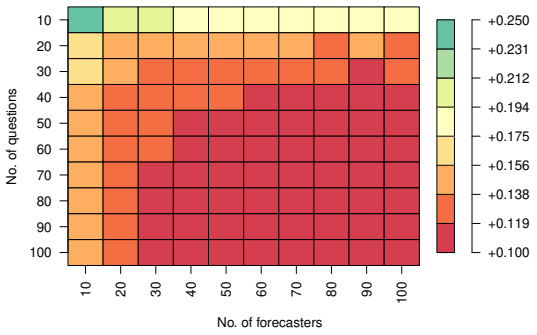
(a) Average Brier scores of E-mean under GJP data.



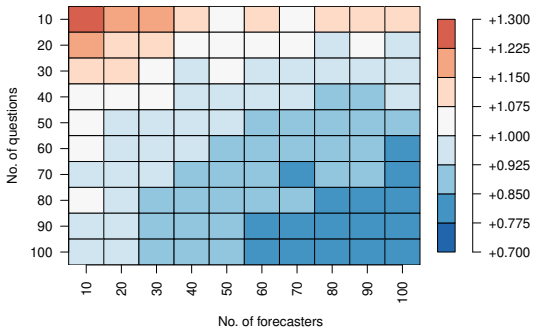
(b) Average Brier scores of E-mean under GK data.



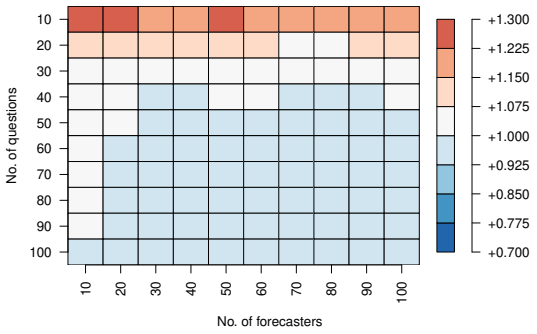
(c) Average Brier scores of SkE-mean under GJP data.



(d) Average Brier scores of SkE-mean under GK data.



(e) SkE-mean Brier scores as a fraction of E-mean Brier scores under GJP data.



(f) SkE-mean Brier scores as a fraction of E-mean Brier scores under GK data.

Figure 7. Out-of-sample average Brier scores of E-mean and SkE-mean aggregators given differing quantities of training data. The left column represents GJP data, and the right column represent GK data. Plots in the same row do not have the same scale, except in Figures 7e and 7f that the present the relative Brier scores. In the bottom row, values smaller (larger) than 1.0 show when SkE-mean performs better (worse, respectively) than E-mean.