Judgment and Decision Making, Vol. 17, No. 1, January 2022, pp. 91–123

Recalibrating probabilistic forecasts to improve their accuracy

Ying Han* David V. Budescu†

Abstract

The accuracy of human forecasters is often reduced because of incomplete information and cognitive biases that affect the judges. One approach to improve the accuracy of the forecasts is to recalibrate them by means of non-linear transformations that are sensitive to the direction and the magnitude of the biases. Previous work on recalibration has focused on binary forecasts. We propose an extension of this approach by developing an algorithm that uses a single free parameter to recalibrate complete subjective probability distributions. We illustrate the approach with data from the quarterly Survey of Professional Forecasters (SPF) conducted by the European Central Bank (ECB), document the potential benefits of this approach, and show how it can be used in practical applications.

Keywords: forecasting, recalibration, extremization, Brier score, human forecasting, subjective probability distributions

1 Introduction

Most forecasting activities involve the ability to reason under uncertainty and require some level of probabilistic reasoning. This is true if one forecasts single quantities (e.g., the value of the market at the end of the current year, or the number of COVID cases that will be recorded next month in a certain country), but it is especially critical when generating probabilistic forecasts. Limited cognitive ability, lack of complete and/or fully reliable information and suboptimal processing of the information available can lead to the pervasive miscalibration in estimating the target probabilities (e.g., Kahneman et al. 1982; Gilovich et al. 2002). There is substantial empirical evidence that judges are often miscalibrated due

*Department of Psychology, Fordham University. Email: yhan23@fordham.edu. https://orcid.org/0000-0001-5289-8527.

†Department of Psychology, Fordham University. Email: budescu@fordham.edu. https://orcid.org/0000-0001-9613-0317.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

to systematic biases (Lichtensten et al. 1982; Zhang & Maloney, 2011) and random errors (Erev, Wallsten & Budescu, 1994). A large number of empirical studies have found that both single point probabilities and probability interval estimates tend to be mostly overconfident (e.g., Alpert & Raiffa, 1982; Budescu & Du, 2007; Juslin, Wennerholm & Olsson, 1999; McKenzie, Liersch & Yaniv, 2008; Fischhoff, Slovi & Lichtenstein, 1977; Klayman et al., 1999; Park & Budescu, 2015). Sometimes miscalibration carries over to specific expertise domains (Christensen-Szalanski & Bushyhead, 1981; Du & Budescu, 2018).

More specifically judges tend to overestimate the probability of, and overweight, rare events and underestimate and underweight highly probable events (Camerer & Ho, 1994; Fischhoff, Slovic & Lichtenstein, 1977; Moore & Healy, 2008; Wu & Gonzalez, 1996) and to avoid extreme probabilities, close to 0 or 1 (Juslin, Winman & Olsson, 2000). Ariely et al. (2000) and Turner et al. (2014) have shown that this tendency of avoiding extreme probability prediction can carry over to aggregated probability estimates.

Baron et al. (2014) attributed the lack of extremity in probability forecasting to two distorting factors. The first, which they labeled an end-of-scale effect, is that the distribution of the estimates near the true value (1 or 0) is not symmetric. Typically, the distribution is regressive towards 0.5, leading to over- (under-) estimation of low (high) probabilities (see analysis in Erev, Wallsten & Budescu, 1994). This causes forecasters[1] to provide less extreme estimates when the true probability is close to the two endpoints (0 and 1). The second factor driving this bias is the forecasters' tendency to mix individual confidence with confidence in the best forecast. Baron et al. (2014) proposed that the extent of reduction in the forecasting extremity is associated with the amount of information that the judge feels is missing.

One natural solution to the problem of miscalibration is to "debias" judges and train them to be better calibrated, but this has turned out to be difficult (Alpert & Raiffa, 1982; Koriat, Lichtenstein & Fischhoff, 1980; Schall, Doll & Mohnen, 2017) and, often, impractical but there are some success stories (e.g., Mellers et al., 2014). An alternative solution, which the subject of this paper, is to recalibrate the judgements (e.g., Shlomi & Wallsten, 2010), i.e., to transform the empirical probability estimates to improve their accuracy. This is a drastically different approach, because the application of these non-linear transformations does not involve the judge(s): They are applied by the users of the estimates or by intermediaries (e.g., decision analysts) before using the forecasts to make actual decisions. For example, if a Decision Maker (DM) makes periodical decisions regarding his/her investment portfolio and believes that his/her financial advisors are systematically biased, he/she may recalibrate the estimates hoping to reduce, if not fully eliminate, this bias before making his/her decisions.

Various transformation methods have been developed and proved to enhance the forecasting accuracy (Ariely et al., 2000; Baron et al., 2014; Satopää et al., 2014; Turner et al., 2014; Mandel, Karvetski & Dhami, 2018). Turner et al. (2014) discussed the Lin-

[1]We use the terms *judge, forecaster* and *expert* interchangeably throughout the paper.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

ear Log Odds (LLO) recalibration function, which has been widely used to compensate the distortion of individual probability forecasts (Gonzalez & Wu, 1999; Tversky & Fox, 1995). The LLO transformation recalibrates the original probability p, by means of a linear transformation of the original log odds, to obtain the recalibrated value, $\widehat{p}$:

$$\widehat{p} = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}. \tag{1}$$

This formula is derived from the linear log-odds model:

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = \gamma \log\left(\frac{p}{1-p}\right) + \tau \tag{2}$$

where γ represents the slope and τ represents the intercept and $\delta = \exp(\tau)$ in Equation 1. Turner et al. (2014) interpreted γ as discriminability parameter which is manifested as curvature of the LLO function. More specifically, when γ increases (decreases), the curve becomes steeper (flatter) in the middle of the range. The other parameter δ was interpreted as the overall response tendency parameter, representing the vertical distance of the curve from zero.

The LLO function can be simplified by restrictions of its parameters to generate special cases of the general family of transformations. When $\delta = 1$ and $\gamma = 1$, $\widehat{p} = p$, the function represents no transformation, and when $\delta = 1$, LLO function becomes the well-known Karmarkar equation (Karmarkar, 1978):

$$\widehat{p} = \frac{p^{\gamma}}{p^{\gamma} + (1-p)^{\gamma}} \tag{3}$$

This function has some attractive properties: (1) it generates probabilities (and does not require any additional normalizations) for binary events, for any value of γ; (2) $\widehat{p} = p$ for three "natural" anchor points $p = 0$, 0.5 and 1. The full LLO function and its simplified version (Equation 3) have been applied in a large body of studies and shown to enhance the accuracy of individual forecasts as well as aggregated forecasts (e.g., Atanasov et al., 2017; Budescu et al., 1997; Baron et al., 2014; Erev et al., 1994; Han & Budescu, 2019; Mellers et al., 2014; Satopää & Ungar, 2015; Shlomi & Wallsten, 2010; Turner et al., 2014).

Mellers et al. (2014) applied Karmarkar's transformation to data generated by more than 2,000 forecasters in a geopolitical forecasting tournament (Aggregative Contingent Estimation ACE; https://www.iarpa.gov/research-programs/ace). They showed that recalibration improved the quality of aggregated probability judgments with optimal γ greater than 1 (implying extremization of the original estimates). They also found some cases of de-extremization, with parameters less than 1. Baron et al. (2014) also applied the same transformation function to the dataset of Mellers et al (2014) and demonstrated that extremization can eliminate the two distorting effects (which cause less extremity in aggregated probability forecasts) with different estimated parameters. They also found out that less extremization (smaller γ) is needed for experts than for non-expert groups and median aggregation requires less extremization than mean aggregation.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

Turner et al. (2014) applied the full LLO function to a different group of forecasters who participated in the ACE forecasting tournament. They compared a set of models which varied in terms of whether (1) the transformation was applied before or after the aggregation, (2) the aggregation was applied to original probability forecasts or log odds of forecasts, and (3) hierarchical modeling of individual difference was utilized. They found that a model that first transforms the raw probability estimates and then aggregates them using log odds improves the forecasting quality the most, compared to the simple aggregation and that the hierarchical modeling of individual difference slightly enhances the forecasting quality. A few studies utilized different recalibration methods. Ranjan and Gneiting (2010) applied beta transformation and Satopää et al. (2014) used a logit model to solve the lack of sharpness of probability judgments and improve the accuracy of probability forecasts. In the current paper we focus on the LLO function and seek to extend its use.

We should clarify that there is no single and "best" recalibration approach. The various applications estimate parameters that seek to optimize one aspect of the forecasts, typically their accuracy. Naturally, if various people seek to optimize different features of the forecasts, they may choose different approaches that can lead to different transformations. Previous studies focused on the recalibration of single (point) probability forecasts associated with simple binary events (e.g., What is the probability that it will rain tomorrow in city Z? What is the probability that candidate A will win next month election in country Y?). This is, of course, a widely used elicitation format in forecasting. Yet recent studies have focused on elicitation methods that seek to estimate complete subjective probability distributions of continuous random variables in a relatively efficient way (Abbas et al., 2008; Haran, Moore & Morewedge, 2010; Wallsten, Shlomi, Nataf & Tomlinson, 2016). Abbas et al. (2008) discussed the Fixed Probability (FP) and the Fixed Value (FV) methods, both of which elicit points along the cumulative distribution of a target variable, X. Haran, Moore and Morewedge (2010) formalized and validated the Subjective Probability Interval Estimates (SPIES) in which judges are asked to allocate probabilities to several predefined bins that represent a C-fold (mutually exclusive and exhaustive) partition of the full range of the target variable. Several large-scale forecasting projects including the Survey of Professional Forecasters (SPF) of European Central Bank (ECB; Garcia, 2003) and the Federal Reserve Bank of Philadelphia (Croushore, 1993) utilize this "bin" method to collect expert forecasters' judgments regarding macroeconomic indicators such as inflation and GDP growth rate.

2 The current paper

In this paper, we describe an extension of Karmarkar's transformation function that can be applied simultaneously to any number of points, on the cumulative distribution, $F(X)$[2], of a random variable, X. These $(C-1)$ points on $F(X)$ can be obtained by any of

[2] $F(X)$ denotes the cumulative distribution function of random variable X.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

methods described earlier but, to fix ideas, it is probably best to think of the SPIES (bins) method where judges assign probabilities to each of the C discrete bins. We illustrate this recalibration approach by re-analyzing data from the quarterly Survey of Professional Forecasters (SPF) conducted by the European Central Bank (ECB). We seek to determine under what circumstances can the proposed recalibration method improve the accuracy of the forecasts and what degree of improvement can be expected, and we illustrated how this approach can be used in practice.

2.1 The transformation function

In this context recalibration means moving the distribution away from an uninformed uniform distribution that assigns an equal probability, $1/C$, to each of the C bins. In the binary case the one-parameter Karmarkar function applies a linear transformation to the log-odds of the event. Here we apply the same approach to the ratio of the odds inferred from the probability assigned to any given bin, $(\frac{P(Bin)}{1-P(Bin)})$, to the odds under equal probability (i.e., $1/(C-1)$), so for every bin the recalibrated probability P^* is obtained as:

$$\log\left(\frac{(C-1)\,P^*\,(Bin)}{1-P^*\,(Bin)}\right) = \gamma \log\left(\frac{(C-1)P\,(Bin)}{1-P\,(Bin)}\right) \tag{4}$$

This implies the transformation function:

$$P^*\,(Bin) = \frac{[(C-1)P(Bin)]^\gamma}{[(C-1)P(Bin)]^\gamma + (C-1)\,(1-P\,(Bin))^\gamma} \tag{5}$$

In the binary case, $C = 2$, this formula recovers Karmarkar's transformation (Equation 3). If the parameter $\gamma > 1$, all probabilities $> 1/C$ increase (i.e., move closer to 1) and all probabilities $< 1/C$ decrease (move closer to 0), so the recalibration extremizes the distribution. If the parameter $\gamma < 1$, the pattern is reversed, so the transformation de-extremizes the distribution, and if $\gamma = 1$ the distribution is not transformed. Three "anchor" probabilities – 0, $1/C$ and 1 – are invariant under the transformation for all values of γ. This suggests that for any given C, all transformation curves cross at $p_i = 1/C$. Figure 1 illustrates the effects of the transformation. Figures 1A and 1B apply various parameters to the cases $C = 3$ and $C = 5$, respectively, Figures 1C and 1D display the effect of two transformation ($\gamma = 0.5$ and 2) for various values of C. Naturally, one can use more complex recalibration functions with additional parameters, but we opted to focus on this simple, intuitive and easy to interpret function.

2.2 The calibration function

Whether or not probability forecasts are transformed, ordinal forecasts (such as one implied by the binning of a continuous variables) are assessed by the ordinal Brier score. This scoring function depends on, and is sensitive to, the specific bin that includes the eventual

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

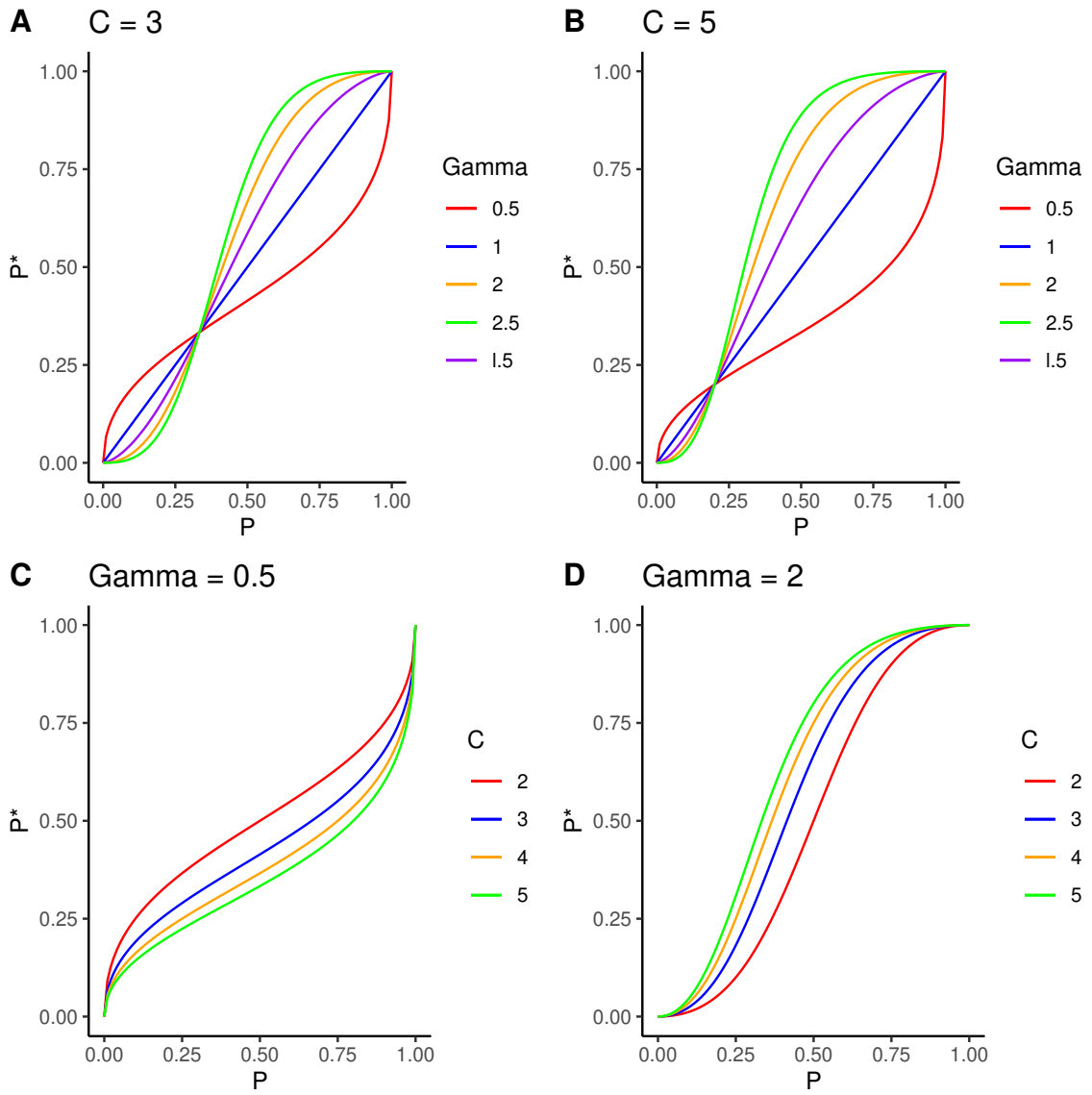


FIGURE 1: The recalibration function for various numbers of bins and transformation parameters.

resolutions of the event. More specifically, if a forecaster assigns the same probability to several bins, it will be scored differentially, as a function of its proximity to the bin of the correct answer. Following Jose et al. (2009) the score is defined by considering all $(C-1)$ binary partitions that preserve the ordering of the categories $[F_1, (1-F_1)]$; $[F_2, (1-F_2)]$; $\cdots$ $[F_{C-1}, (1-F_{C-1})]$[3], (binary) Brier score for each of these partitions and, then, averaging them. If the eventual outcome is in the R'th $(1 \leq R \leq C)$ bin, it is possible to

[3] F_i denotes the cumulative probability of the first i bin(s), i.e., F_2 denotes the cumulative probability of the first 2 bins, specifically, the sum of the first two bin probabilities.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

show that:

$$\text{When } R = 1 : BS = 2[\sum_{i=1}^{C-1}(1 - F_i)^2]/(C-1) \quad (6)$$

$$\text{When } R = C : BS = 2[\sum_{i=1}^{C-1}{F_i}^2]/(C-1) \quad (7)$$

$$\text{Otherwise: } BS = 2[\sum_{i=1}^{R-1}F_i^2 + \sum_{i=R}^{C-1}(1 - F_i)^2]/(C-1) \quad (8)$$

The last formula can be re-expressed in another form that highlights how the BS depends not only on the distribution of forecasts, F_i, but also on the bin of the correct response, R, and the way its location "splits" the distribution over the C bins:

$$BS = 2[(C - R) + \sum_{i=1}^{C-1}F_i^2 - 2\sum_{i=R}^{C-1}F_i]/(C-1) \quad (9)$$

2.3 The recalibration procedure

When recalibrating real forecasts, there are two options for the choice of recalibration parameter γ. One can apply a pre-determined value based on previous experience, experts' advice, etc. Alternatively, one can estimate the optimal parameter γ that maximizes the accuracy of the transformed forecasts where accuracy is measured by the criterion of choice (in our case, the Brier score). We focus on the latter approach and estimate optimal values of γ. The bins (categories) in the ECB data are ordinal, so it is more convenient to recalibrate cumulative probabilities (If we recalibrate specific bin probabilities, we need to add one more step of normalization to make the sum of C recalibrated bin probabilities equal to 1).

With these considerations in mind, we implemented the following recalibration procedure. First, cumulative probabilities F_i for each valid case were computed based on the probabilities assigned to the various bins ($F_i = \sum_1^i P(bin)_i$). Second, the extremization function (Equation 5) was applied to the cumulative probabilities F_i and the recalibration parameter, γ, was estimated by minimizing the corresponding ordinal Brier Score (BS). Finally, the optimal parameter was applied to the relevant probabilistic forecasts and the forecasting performance was evaluated by calculating the ordinal BS of the recalibrated forecasts.

Consider one forecaster in the ECB data set: In 2001Q1, participant (ID # 1) assigned probabilities to the 9 possible bins for the inflation of the current year: {0, 0, 0, 0.15, 0.50, 0.30, 0.05, 0, 0}. The 9 corresponding cumulative probabilities are {0, 0, 0, 0.15, 0.65, 0.95, 1, 1, 1}. Transformation function in Equation 5 was applied to these cumulative probabilities, and transformed probabilities of all 9 cumulative probabilities were expressed as a function including a single parameter γ, i.e., $F_4 = 0.15$, transformed cumulative probability $F_4^* = \frac{(8*0.15)^\gamma}{(8*0.15)^\gamma + 8*0.85^\gamma}$. The ordinal BS was expressed as a function of γ by

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

plugging in transformed cumulative probabilities F_1^* to F_8^* to Equation 8 (the ground truth for this item was 2.35 which is in the 6th bin, hence Equation 8 is appropriate). The optimal parameter γ was estimated by minimizing the ordinal BS function (in this case, $\gamma = 0.62$).[4] The recalibrated cumulative distribution over the $C = 9$ bins after the optimal transformation became {0, 0, 0, 0.13, 0.40, 0.74, 1, 1, 1}. The BS of the optimally recalibrated forecasts is BS= 0.062, compared to BS = 0.112 for the original forecasts. The cumulative distributions before and after transformation of this de-extremization example ($\gamma < 1$) are plotted in the upper 2 panels of Figure 2. Another example of extremization ($\gamma > 1$), based on a different forecaster (ID # 2) for the same event on the same round, is provided in the lower two panels. In this case, the raw cumulative probabilities are {0, 0, 0, 0, 0, 0.7, 1, 1, 1}, and after transformation based on optimal $\gamma = 3.91$, the transformed cumulative probabilities are {0, 0, 0, 0, 0, 0.99, 1, 1, 1}.

3 Data

The Survey of Professional Forecasters (SPF) is a quarterly survey conducted by the European Central Bank (ECB) since 1991. The experts forecast some macroeconomic indicators for the European Union (EU). The experts are affiliated with financial or non-financial institutions in the EU. At each release of the survey, participants are asked to report their forecasts about future HICP (Harmonised Index of Consumer Prices) inflation, the real GDP growth rate and the unemployment rate of in the Euro zone.

The ECB survey elicits expectations for different forecasting horizons (forecasts of the current year, next calendar year, year after next year and five/six[5] years ahead of current calendar year) for the three variables. The survey elicits both point estimates of these quantities as well as full probability distributions of target quantities using the bin method (Haran, Moore & Morewedge, 2010) (see sample questionnaire in Appendix A). We re-analyze the probability distributions for the period starting on the first quarter of 2001 (2001 Q1) to the last quarter of 2017 (2017 Q4), i.e., 72 successive quarters.[6] A total of 99 experts forecasted at least once during this period, but not every expert forecasted all quantities every quarter. Our data set consists of 14,117 forecasts, which translates into an average of about 196 forecasts per target quantity per quarter and 49.02 forecasts per target quantity and time horizon[7] in a quarter. Some cases were removed for the following four reasons (see details in Table 1):

1. Missing cases (no values were assigned to any bins).

[4]This optimization was realized by R function optim() with the optimization algorithm, the adjusted quasi Newton method based on Byrd et al. (1995) with the lower bound of 0.

[5]Expectations of five years ahead of current calendar year are asked in the surveys administered in the first and the second quarters, six years ahead in the Q3 and Q4 surveys.

[6]The data were downloaded from the ECB Statistical Data Warehouse https://sdw.ecb.europa.eu/.

[7]Not all forecast horizons are forecasted at every quarter. Changes in forecast horizon structure over time are documented in Appendix B.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

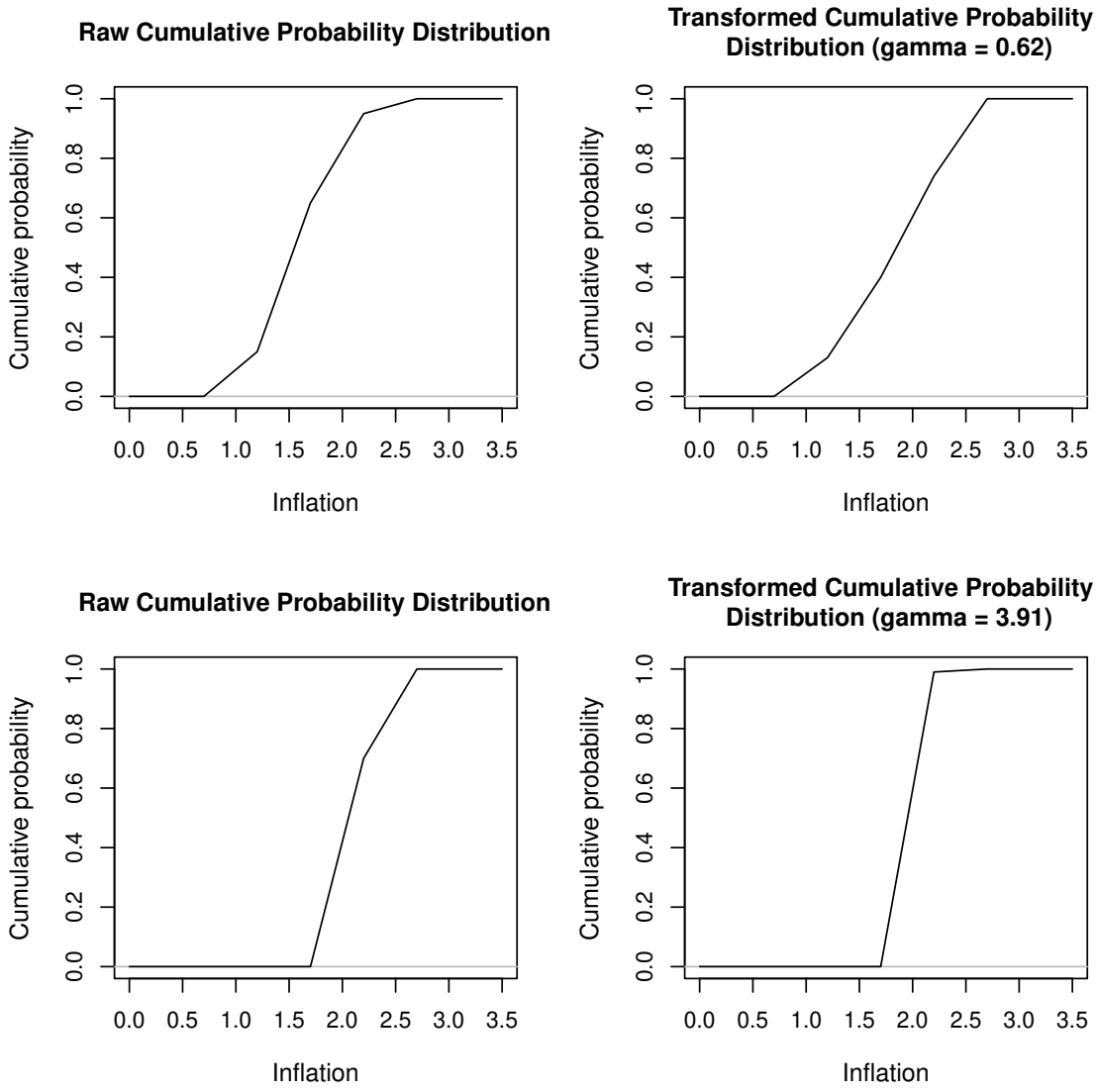


FIGURE 2: Two examples of recalibration using the ECB inflation data.

2. The probability estimates of the C bins did not sum up to 1.
3. True values (ground truths) for the target quantities were not available at the time of the analysis, e.g., probability estimates of year 5 in 2017 Q3 (forecasts of 2022).
4. The parameter estimation procedure failed.

The number of bins and their corresponding upper and lower bounds of all three indicators are determined by the ECB and change over time, as shown in Table 2.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

TABLE 1: Reasons for excluding forecasts from the data set

	Original cases	Forecasts missing	Violating unitarity axiom	No ground truth	Optimization failure	Cases analyzed
Inflation	14117	2473	617	653	2	10371
GDP	14117	2628	584	646	1	10257
Unemployment	14117	3321	527	573	1	9695

TABLE 2: The number of bins and their corresponding ranges for the various indicators.

	Time period	# of bins	Range
Inflation	2001Q1–2008Q2	9	(0, 3.5)
	2008Q3–2009Q1	10	(0, 4)
	2009Q2–2009Q4	14	(–2, 4)
	2010Q1–2017Q4	12	(–1, 4)
GDP	2001Q1–2008Q3	10	(0, 4)
	2008Q4–2009Q1	12	(–1, 4)
	2009Q2–2009Q4	24	(–6, 4)
	2010Q1–2017Q4	12	(–1, 4)
Unemployment	2001Q1–2002Q1	13	(6.5, 12)
	2002Q2–2009Q1	13	(5.5, 11)
	2009Q2–2009Q4	21	(5.5, 15)
	2010Q1–2017Q4	19	(6.5, 15)

4 Results

Basic descriptive statistics of re-calibration parameter and the corresponding BS of three macroeconomic indicators are summarized in Table 3. The cumulative distributions of estimated parameters of three indicators across all the cases are presented in Figure 3. There is a considerable number of values close to 0, and a large number of very high values. It appears that GDP requires the least amount of recalibration, and also a considerable amount of cases are optimized with de-extremization (optimal $\gamma < 1$). The cumulative distribution of the re-calibrated Brier scores is displayed in Figure 4. These scores indicate that GDP is the hardest indicator to predict and inflation is the most predictable. The presence of many high values of parameters distorts the distribution, so we also present (in Figure 5) the distribution of parameters based only on those cases where the estimated parameters do not exceed 10. This eliminates between 10 and 15% of the cases for the various indicators (Table 3).

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

TABLE 3: Descriptive statistics of the recalibration parameter, γ, and the corresponding Brier Scores.

	Variable forecasted	n	$\gamma < 10$	Mean	Median	SD	IQR
Recalibration Parameter, γ	Inflation	10371	0.85	7.111	0.914	17.555	4.076
	GDP	10257	0.87	5.496	1.048	13.405	5.228
	Unemployment	9695	0.9	5.495	0.733	19.947	1.973
Brier Score	Inflation	10371	---	0.105	0.041	0.165	0.128
	GDP	10257	---	0.192	0.069	0.25	0.218
	Unemployment	9695	---	0.122	0.024	0.228	0.119

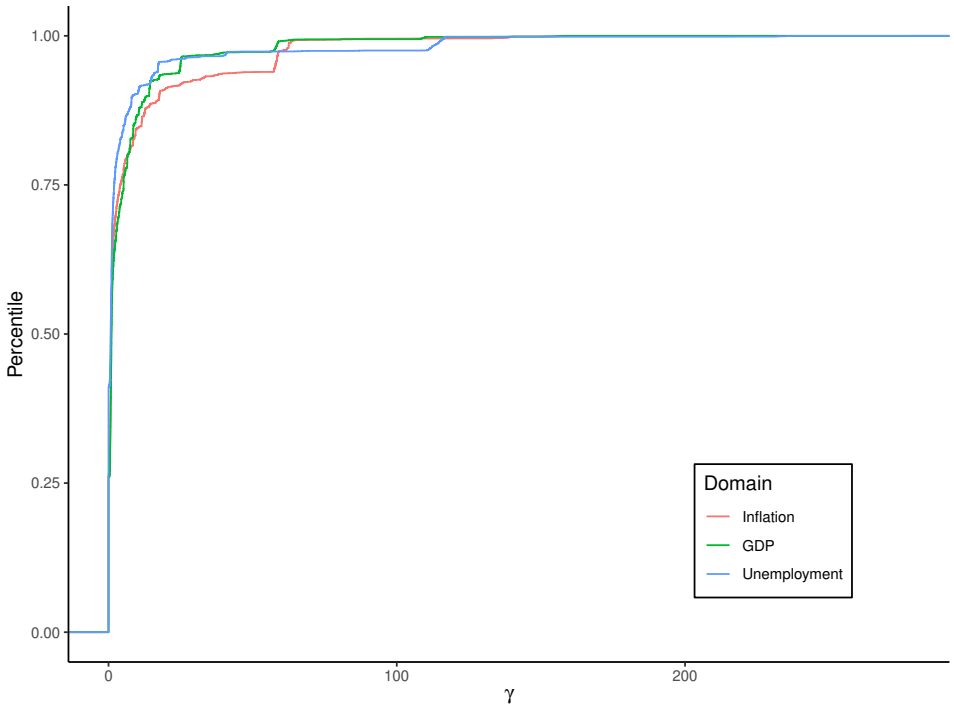


FIGURE 3: Cumulative distributions of the re-calibration parameter (γ) of the three economic indicators.

4.1 Re-calibration parameters and forecasting horizon

Descriptive statistics of re-calibration parameters of different forecasting horizons (FH) of all three economic indicators are summarized in Tables 4–6. We present here only the cases where $\gamma \leq 10$. Analyses of the full data set yield similar results and are relegated to Appendix C.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

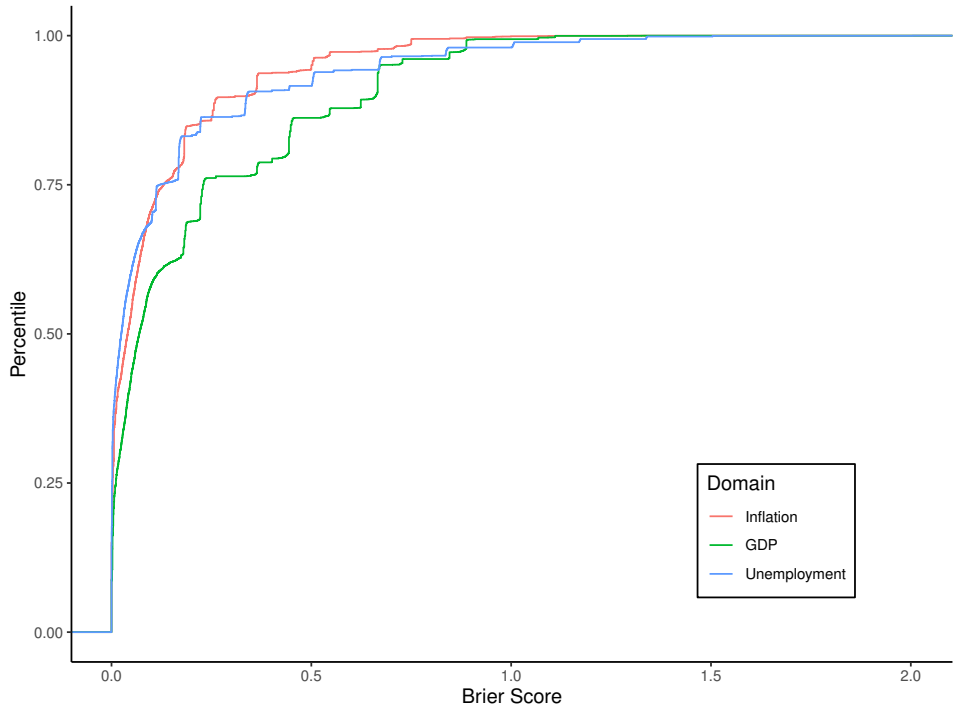


FIGURE 4: Cumulative distributions of the recalibrated Brier Scores of three economic indicators.

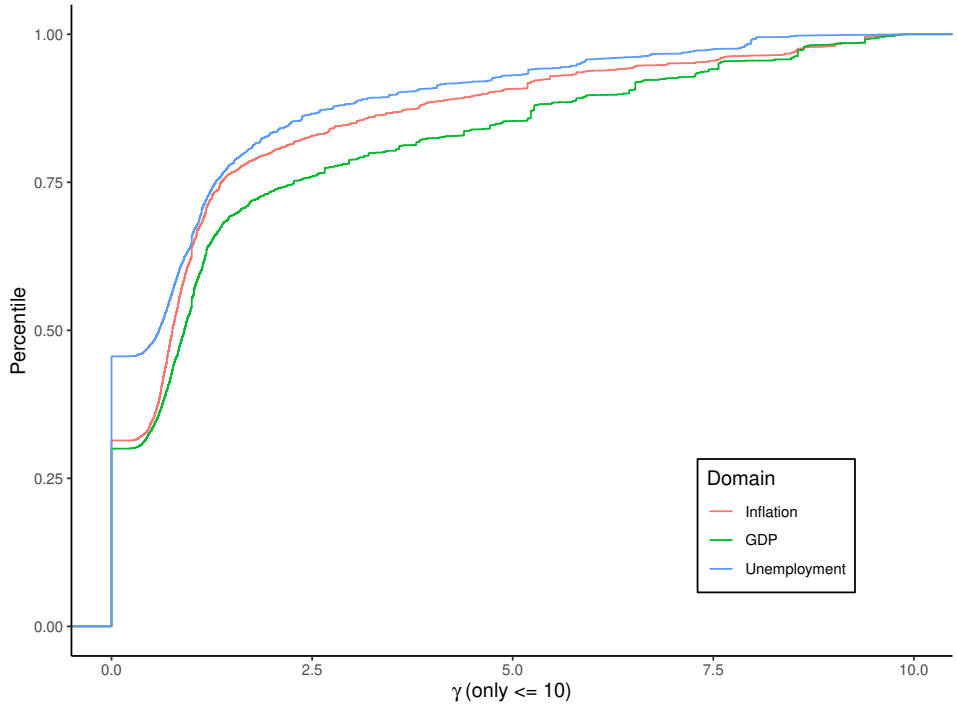


FIGURE 5: Cumulative distributions the re-calibration parameter of the three indicators for cases where $\gamma \leq 10$.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

Several regularities stand out in these displays: (1) There is a general (but not strict[8]) monotonic pattern of the mean parameter values – longer-term forecasting is optimized with higher recalibration parameters compared to shorter forecasting horizons in inflation and GDP. This pattern does not hold for unemployment rate where years 5 and 6 do not yield higher estimated parameters than years 1, 2 and 3; (2) In almost all the cases, the variability of the optimal parameters (as measured by their SDs and IQRs) increases as function of the forecasting horizon (Here, again, the variability in the unemployment rates for years 5 and 6 are unusual); (3) The distributions of the estimated parameters are skewed to the right: In most cases the median γ is below 1, indicating that the majority to the forecasts are being de-extremized to some degree but, on the other hand, a minority of cases induce vary large extremization driving the mean γ.

TABLE 4: Recalibration parameters for Inflation by forecasting horizon (FH) ($\gamma \leq 10$).

FH	n	Mean	Median	Balance	SD	IQR	Skew
The current year	2994	1.43	0.86	–0.18	2.04	1.33	0.28
Next year	2841	1.14	0.63	–0.49	1.88	1.02	0.27
Year after next year	1359	1.7	0.76	–0.27	2.46	1.72	0.38
Year 5/6	1573	1.91	0.98	–0.02	2.42	2.49	0.38

Notes: Balance $= \frac{(\text{cases with } \gamma>1 - \text{cases with } \gamma<1)}{(\text{cases with } \gamma>1 + \text{cases with } \gamma<1)}$.
Skew $= (\text{Mean} - \text{Median})/SD$.

TABLE 5: Recalibration parameters for GDP by forecasting horizon (FH) ($\gamma \leq 10$).

FH	n	Mean	Median	Balance	SD	IQR	Skew
The current year	3059	1.44	0.74	–0.26	2.19	1.39	0.32
Next year	2791	2.06	0.91	–0.09	2.65	3	0.43
Year after next year	1428	1.84	0.91	–0.11	2.45	1.78	0.38
Year 5/6	1614	2.32	1.18	–0.24	2.54	2.91	0.45

Notes: Balance $= \frac{(\text{cases with } \gamma>1 - \text{cases with } \gamma<1)}{(\text{cases with } \gamma>1 + \text{cases with } \gamma<1)}$.
Skew $= (\text{Mean} - \text{Median})/SD$

To confirm that forecasts of longer terms require higher recalibration, we combined the four forecasting horizons into two classes, with Years 1 and 2 representing "short term" and Years 3 and 5/6 representing "long term", and compared parameters of the two classes for all three economic indicators. Table 7 shows that long term forecasts always require larger

[8]The forecasting of next year's GDP yields higher estimated average γ than the forecasting of the year after, which does not follow the general pattern.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

TABLE 6: Recalibration parameters for Unemployment by forecasting horizon (FH) ($\gamma \leq 10$)

FH	n	Mean	Median	Balance	SD	IQR	Skew
The current year	3026	1.09	0.77	–0.24	1.6	1.22	0.2
Next year	2762	1.41	0.67	–0.21	2.1	1.73	0.35
Year after next year	1227	1.68	0.74	–0.15	2.28	2.6	0.41
Year 5/6	1738	0.54	0	–0.7	1.22	0.65	0.44

Notes: Balance = $\frac{(\text{cases with } \gamma>1 - \text{cases with } \gamma<1)}{(\text{cases with } \gamma>1 + \text{cases with } \gamma<1)}$.
Skew = (Mean – Median)/SD

parameters, indicating that estimates of distant events are likely to be more conservative than those of closer events, therefore require greater extremization to optimize the accuracy. This observation is confirmed by the significant t-tests between the two time horizons for inflation ($t(4,899) = –10.2$, $p < .05$) and GDP ($t(6,011) = –6.45$, $p < .05$), but is not supported by the unemployment forecasts, $t(6,072) = 5.57$, $p > .05$.

TABLE 7: Recalibration parameters (γ) for the three indicators for short- and long-term forecasts.

Indicator Forecasted	Forecast Horizon	n	Mean	Median	SD	IQR
Inflation	Short Term	5835	1.288	0.733	1.968	1.208
	Long Term	2932	1.817	0.837	2.439	2.185
GDP	Short Term	5850	1.736	0.828	2.441	2.084
	Long Term	3042	2.094	1.071	2.511	2.147
Unemployment	Short Term	5788	1.241	0.732	1.861	1.386
	Long Term	2965	1.01	0	1.827	1.067

4.2 Brier score improvement and forecasting horizons

In this section we document the benefits of recalibration, in terms of Brier scores. Let Relative Brier Score Difference (RBSD) measure the improvement in accuracy that can be attributed to re-calibration. More specifically, let

$$\text{RBSD} = \frac{\text{Raw BS} - \text{Extremized BS}}{\text{Raw BS}} \tag{10}$$

Higher (lower) RBSD indicates more (less) improvement in the forecasting quality. Overall, recalibration significantly improved the accuracy of the forecasts (Mean RBSD = .569, .452 and .574 for the three indicators). Figures 6–8 summarize the RBSD for different forecasting

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

horizons, showing that distant forecasting horizons yield lower RBSD and benefit less from recalibration, compared to the closer forecasting horizons. In fact, recalibration is most effective and beneficial for year 1 (Mean RBSD = .621, .539 and .623 for the three indicators).

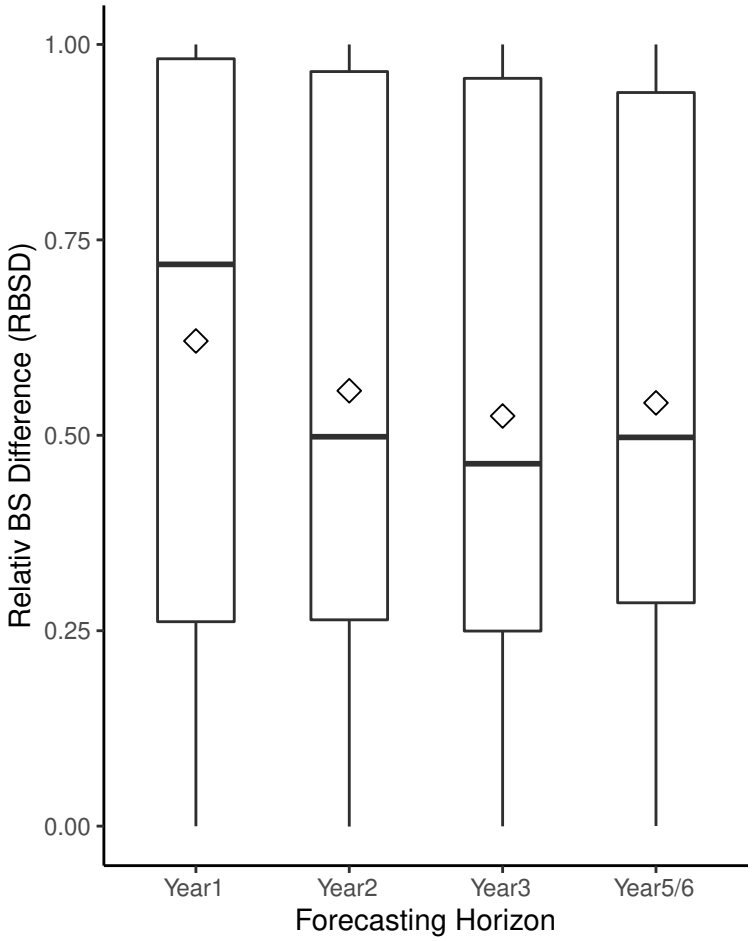


FIGURE 6: RBSD as a function of forecasting horizon (Inflation).

4.3 Practical applications involving out of sample re-calibration

In the previous sections we estimated optimal re-calibration parameters (γ) for each forecast and used these case-specific estimates to re-compute BS and illustrate the effectiveness of the approach. These are, essentially, proofs of concept results but this analysis is analogous to in-sample prediction and, as such, subject to overfitting the data. This is neither a practical approach for predicting future events, nor the optimal method for testing the efficacy of the approach in real-life applications.

In practical settings, one would estimate the optimal parameters based on past performance. This is only possible after the ground truth is revealed which, in the cases studied here, takes a long time. More precisely, the minimal waiting period is the target time

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

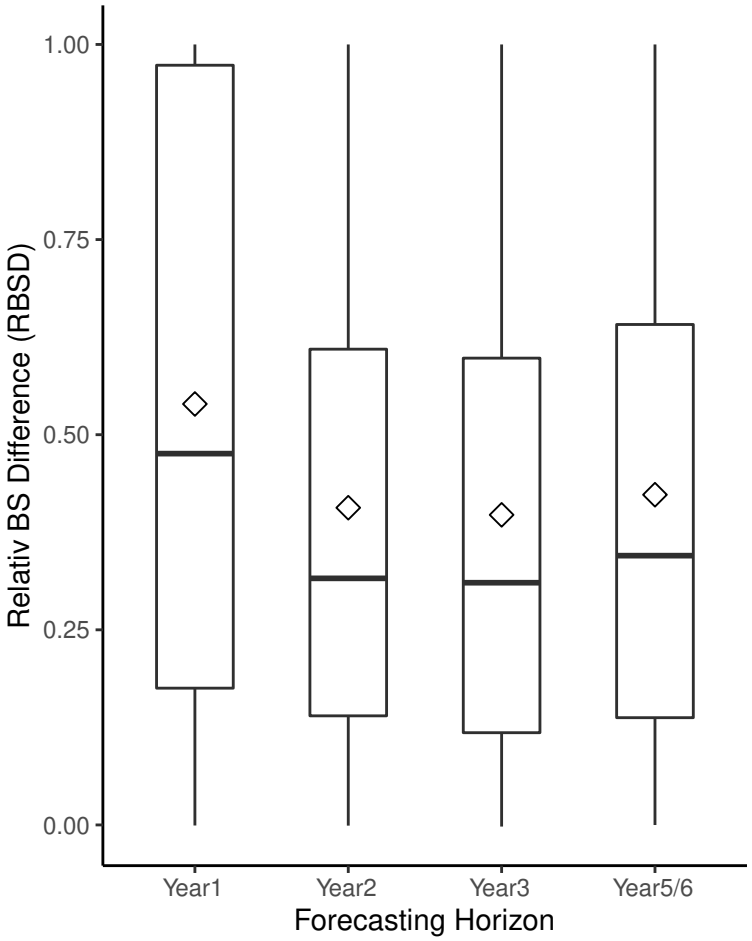


FIGURE 7: RBSD as a function of forecasting horizon (GDP).

horizon. For example, if at time t we wish to predict the value of an indicator at time $(t+k)$ we need to rely on the optimal aggregate of the forecasts provided at time $(t-k)$, which resolve, and allow estimation of γ, only at time t.

Another consideration that affects the best use of historical information is how to best utilize case-specific estimates of the past quarters (the group of judges forecasting in every quarter may also change over time, so it is impossible to generate individual-specific parameters). To explore practical strategies for recalibrating forecasts, we compared the performance of five different types of re-calibration parameters that "borrow" information from other forecasts and forecasters.

1. Domain-specific: The median of all the case-specific parameters (γs) (collapsing all the time horizons and quarters) of any given economic indicator.
2. Quarter-specific: The median of all the case-specific parameters (γs) (collapsing all the time horizons) of the same quarter for each economic indicator.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

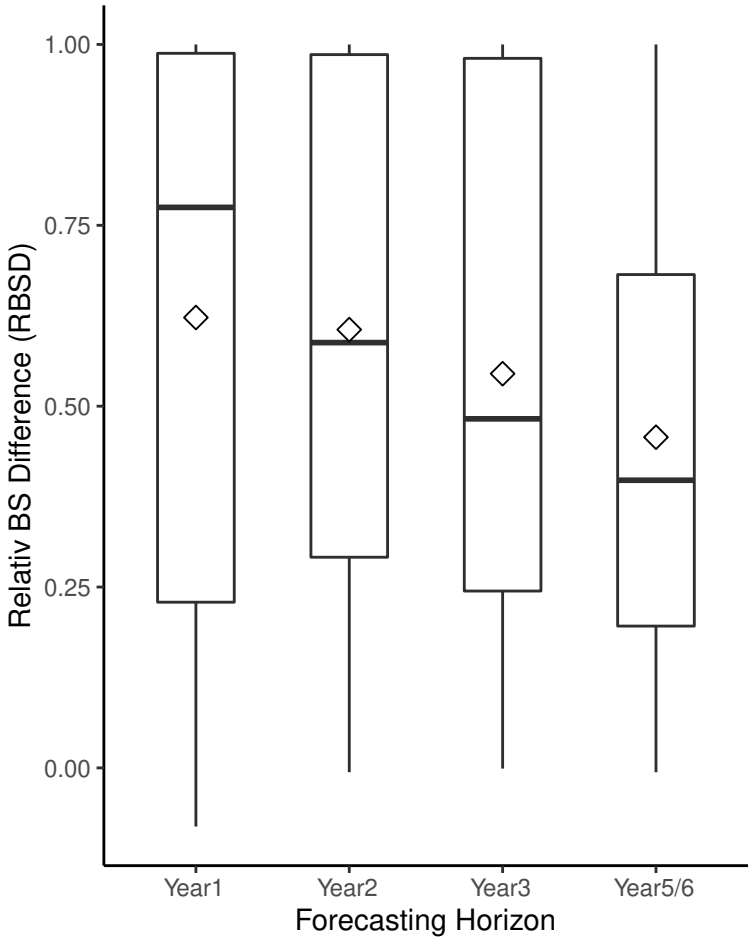


FIGURE 8: RBSD as a function of forecasting horizon (Unemployment).

3. Forecast horizon specific: The median of all the case-specific parameters (γs) (collapsing all the quarters) of the same forecasting horizon for each economic indicator.
4. Quarter & Forecast horizon specific: The median of all the case-specific parameters (γs) of the same forecasting horizon and the same quarter for each economic indicator.
5. Aggregate: Estimate the optimal parameter (γ) for the mean probability distribution[9] for any given quarter and FH for every indicator.

We calculated the parameters based on these five approaches, used them to re-calibrate the forecasts, and we compared the BS obtained from the different selections to the performance of two baselines: No recalibration ($\gamma = 1$) and the optimal recalibration based on the case-specific γ. The results are summarized in Table 8. For all three indicators, the aggregate γ performs best (closest to the case specific upper bound) and the quarter &

[9]The mean aggregated parameter is computed in two steps: First, the aggregated forecast is obtained by taking the average of assigned probabilities of the same bin across all judges; Second, the optimization procedure is applied to obtain the optimal extremization parameter for the aggregated forecast.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

forecasting horizon specific γ performs the second best. Both approaches systematically outperform the original (untransformed) forecasts across all three domains.

TABLE 8: Brier scores of different types of recalibration parameters for three indicators.

	Inflation		GDP		Unemployment	
	Mean	Median	Mean	Median	Mean	Median
No extremization	0.217	0.14	0.299	0.18	0.219	0.109
Domain	0.2	0.123	0.3	0.184	0.204	0.111
Quarterly	0.179	0.1	0.367	0.252	0.184	0.102
FH	0.2	0.119	0.3	0.179	0.205	0.109
Quarterly & FH	0.141	0.076	0.219	0.11	0.159	0.084
Aggregated BS	0.092	0.064	0.15	0.063	0.046	0.032
Case-specific	0.105	0.041	0.192	0.069	0.122	0.024

Note: The best two methods are highlighted

Given these results, we estimated first the two top-performing γ parameters (the optimal aggregate γ and the quarter & forecasting horizon specific γ) in every quarter and for every relevant time horizon for the three indicators and used these parameters to re-calibrate the relevant forecasts (i.e., same time horizon for each indicator) for the next period. For example, if forecasts made at 2002Q1 target one calendar year ahead, we estimated the best γ based on forecasts made a year earlier (2001Q1) as soon as the target events resolved (at 2002Q1) and used them to predict the next round of forecasts for the same time horizon (2003Q1).

Table 9 presents the mean Brier scores across all relevant quarters for every time horizon and indicator. The first and the second panels of the table show the original Brier scores (untransformed) and the case-specific Brier scores as the lower and the upper benchmarks. The third panel shows the Brier scores based on the recalibrated forecasts based on the optimal aggregate γ of the previous period and the fourth panel shows the scores based on the quarter & forecasting horizon specific γ of the previous period. Both sets of γ parameters estimated from the previous periods outperform the untransformed BS, but only for short-term forecasts for the current year. For the longer horizons the performance of the optimal parameters based on the previous periods does not outperform the baseline BS. This pattern is consistent for all three indicators.

Table 10 focuses on the current year forecasts for the various indicators and displays the number of *individual forecasts,* where applying one of the two approaches improved or, conversely, caused a deterioration in the Brier score (we excluded cases where $\gamma = 1$, and the Brier score in unaltered.). In a significant majority of the cases, recalibration using the two top-performing estimates of the γ parameter from the previous period was

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

TABLE 9: Performance of the optimal aggregated γ and the quarter & forecasting horizon specific γ of the previous time period.

	Economic Indicator		
Time Horizon	Inflation	GDP	Unemployment
Mean Original Brier Scores (γ=1)			
Current year	0.09	0.186	0.071
Next year	0.255	0.365	0.168
Year after next year	0.275	0.32	0.263
Year 5/6	0.32	0.364	0.527
Mean Case-specific Brier Scores			
Current year	0.033	0.101	0.027
Next year	0.128	0.245	0.083
Year after next year	0.14	0.208	0.151
m Year 5/6	0.16	0.243	0.33
Mean Recalibrated Brier Scores Using Aggregated BS γs Estimated in Previous Period			
Current year	0.061	0.127	0.048
Next year	0.252	0.406	0.175
Year after next year	0.365	0.472	0.305
Year 5/6	0.333	0.369	0.594
Mean Recalibrated Brier Scores Using Quarterly & FH γs Estimated in Previous Period			
Current year	0.078	0.131	0.061
Next year	0.283	0.448	0.199
Year after next year	0.394	0.496	0.345
Year 5/6	0.428	0.385	0.588

Note: Cases where recalibration improved the Brier Scores are highlighted.

successful. Optimal aggregated γ of previous period yields better (lower) BSs compared to the untransformed baseline in at least 77% cases for the three indicators. Quarter & horizon specific γ's of previous period improved the BSs in at least 69% for the three indicators.

The two sets of estimates of quarter and domain specific γs are highly consistent, as shown in Figure 9. After excluding a few extreme estimates, and concentrating only on cases where $\gamma \leq 10$, the two sets correlate highly ($r = 0.87$). The aggregated γ performed better than quarter & horizon specific γ and the out-sample recalibration worked best for the GDP forecasts, for both methods.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

TABLE 10: Distribution of short term individual forecasts where recalibrations improved accuracy

	Economic Indicator			
	Inflation	GDP	Unemployment	Total
	Aggregated BS γ's Estimated in Previous Period			
Number of forecasts	3259	3281	3127	9667
Better than baseline	2,497 (76.6%)	2,686 (81.9%)	2,426 (77.6%)	7,609 (78.7%)
Worse than baseline	762 (23.4%)	595(18.1%)	701 (22.4%)	2,058 (21.3%)
Balance	0.53	0.64	0.55	0.57
	Quarterly & FH γ's Estimated in Previous Period			
Number of forecasts	3211	3281	3083	9575
Better than baseline	2,222 (70%)	2,535 (77.3%)	2,115 (68.6%)	6,872 (71.8%)
Worse than baseline	989 (30%)	746 (22.7%)	968 (32.4%)	2,703 (28.2%)
Balance	0.38	0.55	0.37	0.44

Note: Balance $= \frac{\text{Better}-\text{Worse}}{\text{Better}+\text{Worse}}$.

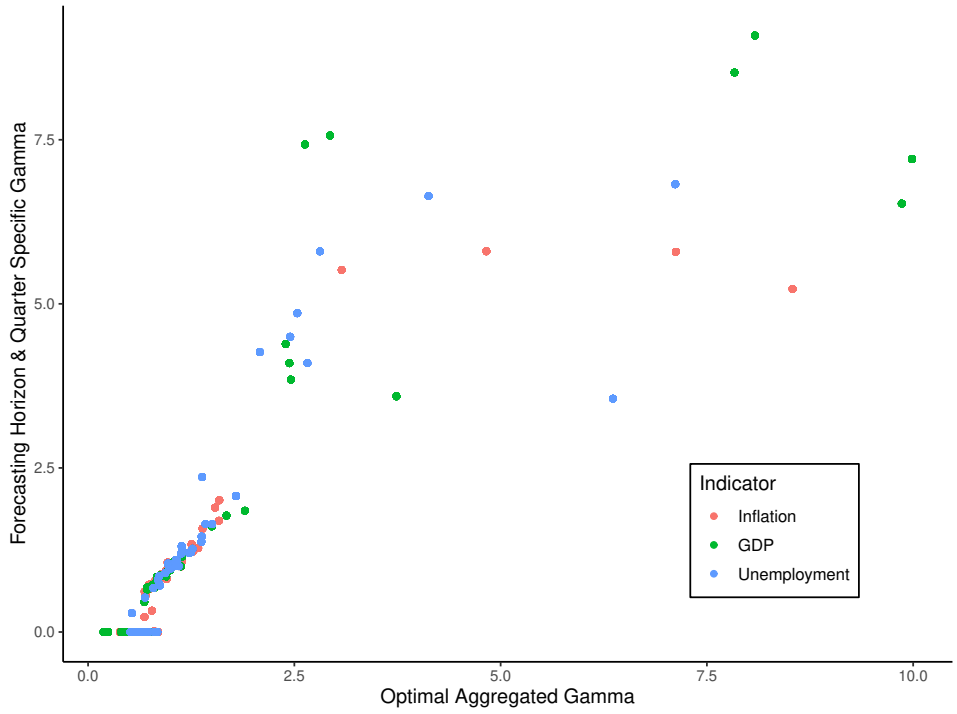


FIGURE 9: Joint distribution of two sets of out-of-sample, quarter and domain specific, estimates of the recalibration parameters (all γs $\leq$ 10).

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

5 Beyond Brier Scores

Our approach was driven by the desire to improve the accuracy of the probabilistic forecasts, as measured by their Brier Scores. This choice is motivated and justified by the fact that accuracy is, typically, the top desideratum of good forecasts, and that the Brier Scores are considered by many the "gold standard". For example, they are often used in forecasting competitions (e.g., Himmelstein, Atanasov & Budescu, 2021; Mellers et al., 2014). However, as some of the reviewers of this manuscript have pointed out, this is not the sole criterion one could consider and, in fact, several appealing alternatives are well documented (e.g., Steyvers, Wallsten, Merkle & Turner, 2014).

In this section we illustrate the effect of the recalibration on an alternative quality measure. Many people prefer evaluating the quality of forecasts by comparing a single best value, extracted from the distribution, to the ground truth. This approach is seen as simpler and easier to interpret, because its scale is more intuitive than Brier. In this spirit, we calculated the median of each distribution in our sample (Raw and Transformed form) and calculated its Relative Absolute Distance (RAD) to the ground truth:

$$RAD = \frac{|\text{Median Estimate} - \text{Ground Truth}|}{|\text{Ground Truth}|}$$

Figures 10–12 display the joint distributions of the Raw and Transformed RADs for the three indicators. Most of the points lie below the respective diagonals indicating that the recalibrated distributions provide more accurate predictions. Thus, on average, and in most individual cases the medians inferred from the recalibrated distributions are closer to the eventual outcomes. The proportion of cases where the recalibration improved the point prediction is 76.02% (Mean improvement = 0.65, SD = 1.49) for Inflation, 78.16% (Mean improvement = 0.48, SD = 1.01) for GDP and 72.23% (Mean improvement = 0.04, SD = 0.07) for Unemployment.

We should clarify that each quality criterion can, in principle, be used to derive an optimal transformation (e.g., one could seek to derive distributions such that their RAD, or other metrics, be minimized). We focused on the Brier score but this example illustrates that this transformation can also benefit other relevant measures of quality.

6 Concluding remarks

There are several compelling examples in the forecasting literature (e.g., Baron et al., 2014; Turner at al., 2014) illustrating the benefits of recalibration of individual forecasts, as well as aggregates of multiple forecasts, of the target events. These examples involve binary events and, as such, amount to recalibrating – extremizing or de-extremizing – a single probability. In this paper we proposed, to our knowledge, the first extension of this approach that allows one to recalibrate a cumulative probability function based on C of its quantiles in a consistent and coherent way that is captured by its single parameter, γ. The recalibration function

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

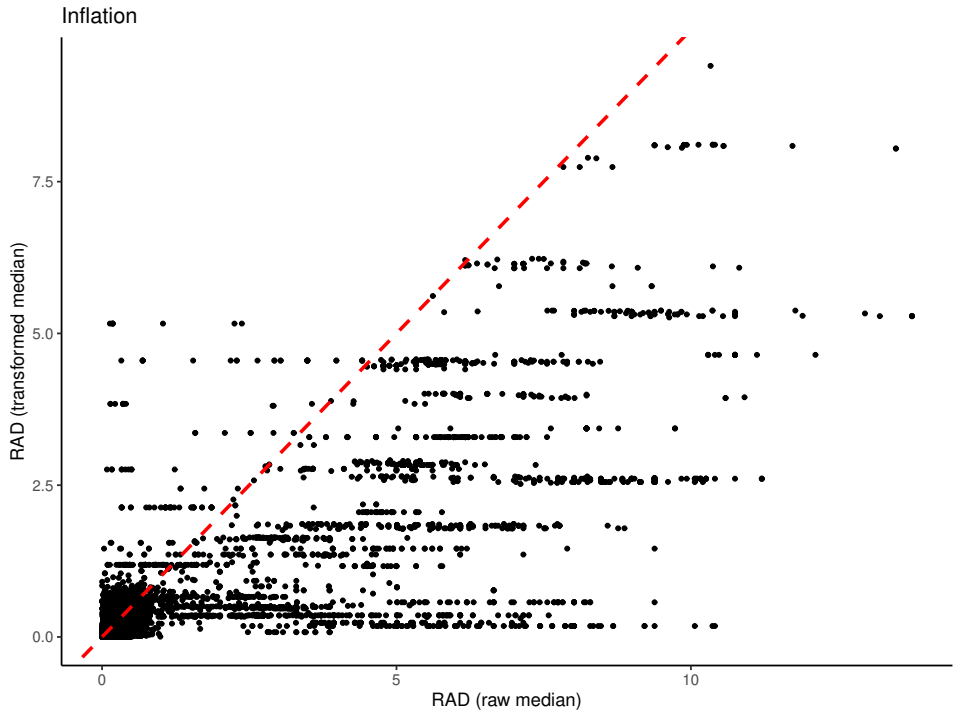


FIGURE 10: Joint distribution of the Raw and Transformed RADs for Inflation.

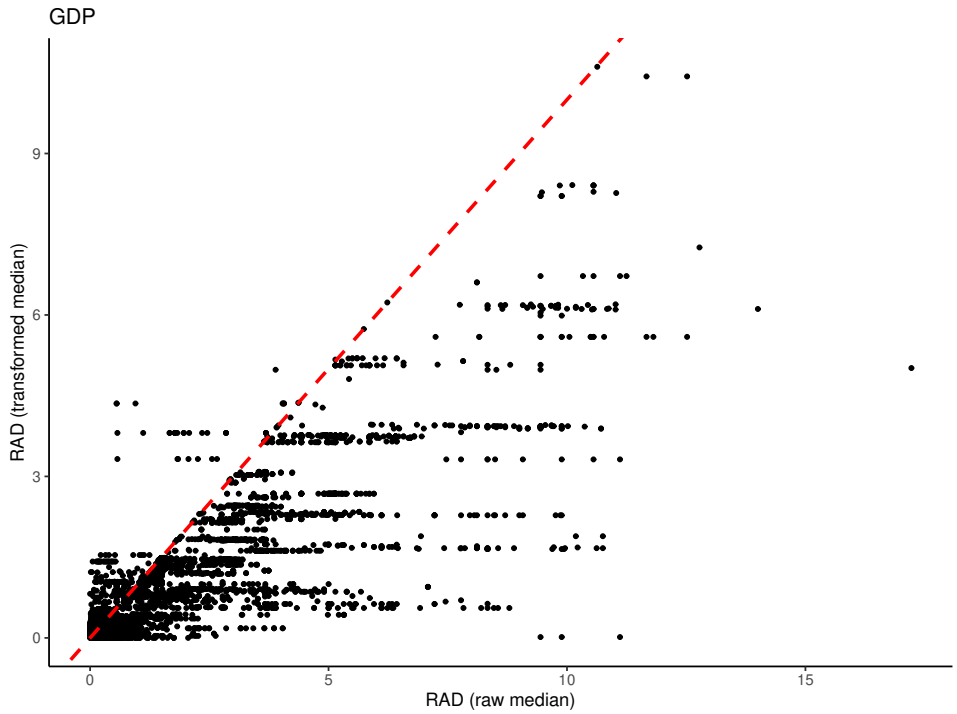


FIGURE 11: Joint distribution of the Raw and Transformed RADs for GDP.

is defined relative to the uniform distribution and its impact is defined in relation to the invariant "anchor", Prob = $1/C$, in the sense that probabilities below or above this anchor

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

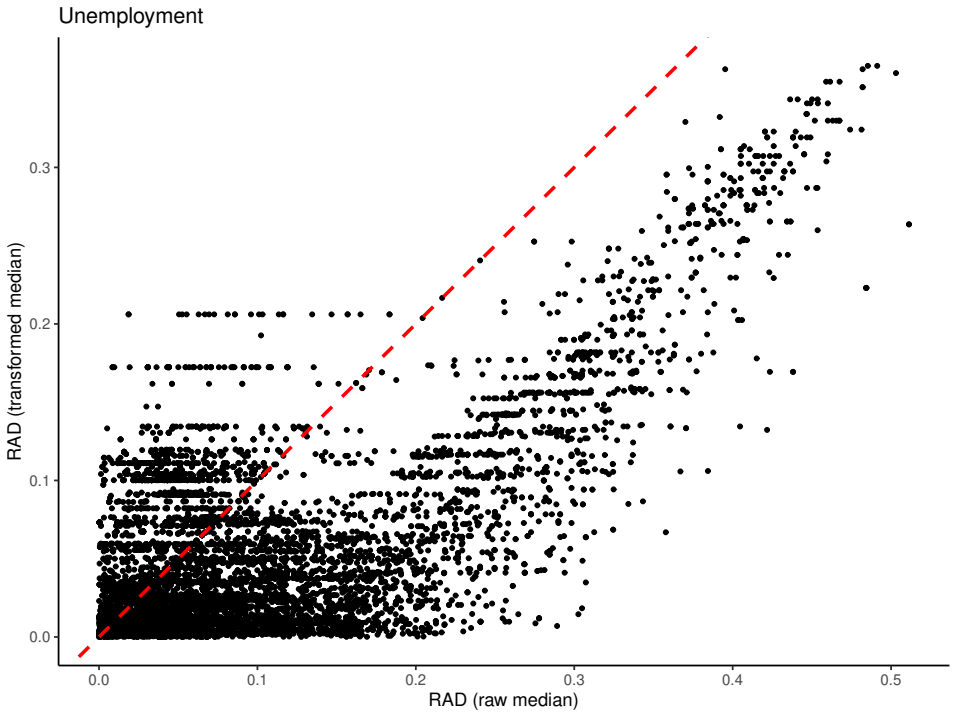


FIGURE 12: Joint distribution of the Raw and Transformed RADs for Unemployment.

are transformed in different directions. The recalibration function generalizes Karmarkar's transformation that was used often in the special case $C = 2$.

We discussed some of the properties of the proposed function and illustrated its use by re-analyzing a large body of forecasts for three economic indicators made by almost 100 experts and spanning 72 quarters. This analysis confirmed that recalibration can be highly beneficial (see Figures 6–8) and we found that its effects are not uniform, in the sense that not all indicators benefit equally. It also clearly showed that, on average, longer term forecasts require more aggressive recalibration. Finally, we have illustrated obvious practical applications of our approach by showing how one can use recalibration parameters estimated in previous periods to significantly increase the accuracy of future short-term forecasts.

We make no claims of optimality or uniqueness regarding our approach. The method we used was develop as a straightforward generalization of the simplest function used in binary cases, using a single parameter. We expect that more complex function could improve accuracy further, and we hope that future work in this area will explore alternative, possibly more flexible and powerful, recalibration functions. One issue we did not study is how the function operates when applied to distributions that are elicited at various levels of precision (i.e., number of bins). In our dataset, the experts were typically given more than 10 bins (see details in Table 2), and we observed that many tail bins were often assigned probabilities of 0. One way to improve the recalibration process may be to develop algorithms that are sensitive to the total number of bins and/or the way the judges use them.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

An interesting question that was raised by one of the reviewers of the paper is whether one should consider forecast recalibration as a one-shot adjustment, or as an additional component to be implemented periodically as part of the forecasting process? We believe that the answer is somewhere in between these two extremes. In a perfectly stable and stationary world, once a transformation function is identified it could be applied routinely to all new forecasts in the same domain. However, recalibration is not perfect (see our results), the estimation is susceptible to random errors and capitalization of chance and, at least in principle, it could be improved as more data become available. And, of course, the world is not stationary and the circumstances that drive the behavior of the target variables of interest, may change over time making older recalibration parameters suboptimal or obsolete.

References

Abbas, A. E., Budescu, D. V., Yu, HT., & Haggerty, R. (2008). A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Analysis*, *5*(4), 190–202. http://dx.doi.org/10.1287/deca.1080.0126.

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P.Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). Cambridge, England: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477.022.

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., & ... Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130–147. http://dx.doi.org/10.1037/1076-898X.6.2.130.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706. http://dx.doi.org/10.1287/mnsc.2015.2374.

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. http://dx.doi.org/10.1287/deca.2014.0293.

Budescu, D.V., Wallsten, T.S. & Au, W. (1997). On the importance of random error in the study of probability judgment. Part II: Using the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making, 10*, 173–188. http://dx.doi.org/10.1002/(SICI)1099-0771(199709)10:3<173::AID-BDM261>3.0.CO;2-6.

Budescu, D.V. & Du, N. (2007). The coherence and consistency of investors' probability judgments. *Management Science*, 53, 1731–1744. http://dx.doi.org/10.1287/mnsc.1070.0727.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization, *SIAM J. Optim*, 9, 877–1208. http://dx.doi.org/0.1137/0916069.

Camerer, C. F., & Ho, T. H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*, *8*, 167–196. http://dx.doi.org/10.1007/BF01065371.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology. Human Perception and Performance*, *7*, 928–935. http://dx.doi.org/10.1037/0096-1523.7.4.928.

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters. *Business Review — Federal Reserve Bank of Philadelphia 3*, November/ December, 3–15. (Retrieved from https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters.

Du, N. & Budescu, D.V. (2018). How (over) confident are financial analysts? *Journal of Behavioral Finance, 19*(3), 308–318. http://dx.doi.org/10.1080/15427560.2018.1405004.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527. http://dx.doi.org/10.1037/0033-295X.101.3.519.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552–564. http://dx.doi.org/10.1037/0096-1523.3.4.552.

Garcia, J. A. (2003). An introduction to the ECB's Survey of Professional Forecasters. Occasional Paper Series, No 8. Frankfurt am Main, Germany: European Central Bank. (Retrieved from https://www.ecb.europa.eu/pub/pdf/scpops/ecbocp8.pdf?b632908ccefcd886a379f074ab6ad12d).

Gilovich, T., Griffin, D., Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge University Press, Cambridge, UK. http://dx.doi.org/10.1017/cbo9780511808098.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*, 129–166. http://dx.doi.org/10.1006/cogp.1998.0710.

Han, Y. & Budescu, D.V. (2019). A universal method for evaluating the quality of aggregators. *Judgment and Decision Making*, *14*(4), 395–411.

Haran, U., & Morewedge, C., & Moore, D. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*(7), 467–476. http://dx.doi.org/10.1037/e615882011-200.

Himmelstein, M., Atanasov, P. & Budescu, D.V. (2021). Forecasting forecaster accuracy: Contributions of past performance and individual differences. *Judgment and Decision Making.16*(2), 323–362.

Jose, V. R., & Winkler, R. L. (2009). Evaluating quantile assessments. *Operations*

Research, *57*(5), 1287–1297. http://dx.doi.org/10.1287/opre.1080.0665.

Juslin, P., & Wennerholm, P & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology. Learning, Memory & Cognition*, *25*(4), 1038–1052. http://dx.doi.org/10.1037/0278-7393.25.4.1038.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*(2), 384–396. http://dx.doi.org/10.1037/0033-295X.107.2.384.

Kahneman, D., Slovic, P., Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases.* Cambridge University Press, Cambridge, UK. http://dx.doi.org/10.1017/CBO9780511809477.002.

Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance*, *21*, 61–72. http://dx.doi.org/10.1016/0030-5073(78)90039-9.

Klayman, J., Soll, J., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247. http://dx.doi.org/10.1006/obhd.1999.2847.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* 6(2), 107–118. http://dx.doi.org/10.1037/0278-7393.6.2.107.

Lichtenstein, S., Fischhoff B., & Pillips, D. (1982). Calibration of probabilities: The state of the art to 1980. In Kahneman, D., Slovic, P. & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 306–334). Cambridge University Press, Cambridge, UK. http://dx.doi.org/10.1017/CBO9780511809477.002.

Mandel, D. R., Karvetski, C. W., & Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making*, *13*(6), 607–621.

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*, 179–191. http://dx.doi.org/10.1016/j.obhdp.2008.02.007.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. http://dx.doi.org/10.1177/0956797614524255.

Moore, D. A. & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. http://dx.doi.org/10.1037/0033-295X.115.2.502.

Park, S. & Budescu, D.V. (2015). Aggregating multiple probability intervals to improve calibration. *Judgment and Decision Making, 10*(2), 130–143.

Ranjan, R. & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *72*(1), 71–91. http://dx.doi.org/10.1111/j.1467-9868.2009.00726.x.

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. http://dx.doi.org/10.1016/j.ijforecast.2013.09.009.

Satopää, V., & Ungar, L. (2015). Combining and extremizing real-valued forecasts. https://arxiv.org/abs/1506.06405.

Schall, D. L., Doll, D., & Mohnen, A. (2017). Caution! Warnings as aword useless countermeasure to reduce overconfidence? An experimental evaluation in light of enhanced and dynamic warning designs. *Journal of Behavioral Decision Making*, *30*(2), 347–358. http://dx.doi.org/10.1002/bdm.1946.

Shlomi, Y., & Wallsten, T. S. (2010). Subjective recalibration of advisors' probability estimates. *Psychonomic Bulletin and Review*, *17*(4), 492–498. http://dx.doi.org/10.3758/PBR.17.4.492.

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.

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. http://dx.doi.org/10.1007/s10994-013-5401-4.

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, *102*, 269–283. http://dx.doi.org/10.1017/cbo9780511803475.006.

Wallsten, T. S., Shlomi, Y., Nataf, C., & Tomlinson, T. (2016). Efficiently encoding and modeling subjective probability distributions for quantitative variables. *Decision*, *3*(3), 169–189. http://dx.doi.org/10.1037/dec0000047.

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, *42*(12), 1676–1690. http://dx.doi.org/10.1287/mnsc.42.12.1676.

Zhang, H., & Maloney, L. T. (2011). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1–14. http://dx.doi.org/10.3389/fnins.2012.00001.

Appendix A: SPF sample questionnaire

16 January 2013

Quarterly Survey of Professional Forecasters (SPF)

Attached is the questionnaire for the 2013 Q1 Survey of Professional Forecasters of euro area macroeconomic variables conducted by the European Central Bank. The following paragraphs give some guidance on how to complete it.

The SPF questionnaire asks for a point forecast of what you expect inflation, real GDP growth and unemployment to be over specific time horizons, together with probabilities for different outcomes. You should not feel obliged to assign probabilities to all of the listed outcomes. **Please note that we have extended the forecast horizon of the SPF**, now asking for forecasts of macroeconomic variables two calendar years ahead already from the first round of the year (they were previously surveyed in the Q3 and Q4 rounds only). Therefore, this questionnaire asks for your expectations **for the calendar years 2013, 2014 and 2015.**

The questionnaire asks as well for your expectations for one and two years ahead of the period for which the most up-to-date official data for each variable are available (you can find these figures below). Specifically in this questionnaire you are asked for your expectations for December 2013 and December 2014 for HICP inflation, for 2013 Q3 and 2014 Q3 for real GDP growth, and for November 2013 and November 2014 for the unemployment rate. Please note that you are also asked for your longer term expectations for inflation (against the ECB Governing Council's aim to keep the annual inflation rate below, but close to, 2% over the medium term), real GDP growth and the unemployment rate. The horizon for these longer term expectations is the year 2017 as a whole.

With regard to inflation and real GDP growth you are asked to fill in your expectations of the year-on-year change in these variables. In the case of the unemployment rate, we would like to know what you expect the level (seasonally adjusted) to be at the specified time horizons. The forecasts should refer to the period averages. We would be thankful if in addition to the quarterly forecasts for oil prices, interest rates and exchange rates, you could also provide **annual average forecasts** of these variables for 2014 and 2015.

There is also a memorandum item included in the questionnaire. Its purpose is to enable you to elaborate on two dimensions: i) any **specific factors that have significantly affected your baseline outlook** for inflation, real GDP growth or unemployment, distinguishing, where possible, between upward and downward revisions compared with last quarter's outlook; ii) **the main risks** associated with the current baseline outlook. **In particular, we would be grateful if you could comment on the point in the forecast horizon in which - according to your baseline outlook – the level of real GDP will increase again in a sustained manner (i.e. witness successive quarter-on-quarter growth rates larger than 0.0%).**

The questionnaire is in Excel format. Please return the completed questionnaire to the ECB **by Tuesday 22 January 2013 cob,** to the e-mail address **ecb-spf@ecb.europa.eu** or the fax number **+49 69 1344-7602**. If you have any questions please contact Nicola Bowen on +49 69 1344-6351, Jeanette Cramer on +49 69 1344-8231, Moritz Karber on +49 69 1344-7222, Victor Lopez Perez on +49 69 1344-5167, Alexandros Melemenidis on +49 69 1344-7179, Asterios Paschos on +49 69 1344-5784 or Coralia Pastora on +49 69 1344-5121.

If your address, telephone number, fax number, e-mail address or contact person (i.e. the person that should receive the survey questionnaire) has changed, please fill in the new information below.

Contact person (job title):
Address:
Telephone:
Fax:
E-mail:

Basic reference data for the 2013 Q1 survey:

- Annual HICP inflation (December 2012): 2.2%
- Annual GDP growth (2012 Q3): -0.6% (according to ESA95)
- Unemployment rate (November 2012): 11.8 %

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

	Point estimate of euro area inflation expectations* Year-on-year change in the HICP					
	2013	**2014**	**2015**	**December 2013**	**December 2014**	**5 years ahead (2017)**
Rate (%)						

* Defined on the basis of the Harmonised Index of Consumer Prices produced by Eurostat.

	Probabilities of euro area inflation* Year-on-year change in the HICP					
	2013	**2014**	**2015**	**December 2013**	**December 2014**	**5 years ahead (2017)**
< -1.0%						
-1.0- -0.6%						
-0.5- -0.1%						
0.0-0.4%						
0.5-0.9%						
1.0-1.4%						
1.5-1.9%						
2.0-2.4%						
2.5-2.9%						
3.0-3.4%						
3.5-3.9%						
≥ 4.0%						
Total	100	100	100	100	100	100

* Defined on the basis of the Harmonised Index of Consumer Prices produced by Eurostat. Probabilities should sum to 100%. Average of the period.

Please report selected other information underlying your forecasts (average over the period):

	2013 Q1	**2013 Q2**	**2013 Q3**	**2013 Q4**	**2014**	**2015**
ECB's interest rate (main refinancing operations)						
Brent crude oil prices (US dollars)						
USD/EUR exchange rate						

	2013	**2014**	**2015**	**2017**
Labour Costs (annual rate of change in whole economy compensation per employee)				

Memorandum item:
In the space below please indicate the main economic factors affecting the outlook for euro area inflation over the forecast horizon distinguishing, where possible, between upward and downward revisions for each year and the main risks associated with this outlook.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

	Point estimate of euro area real GDP growth expectations* Year-on-year change					
	2013	**2014**	**2015**	**2013 Q3**	**2014 Q3**	**5 years ahead (2017)**
Rate (%)						

* Standardised ESA definition.

	Probabilities of euro area real GDP growth* Year-on-year change					
	2013	**2014**	**2015**	**2013 Q3**	**2014 Q3**	**5 years ahead (2017)**
< -1.0%						
-1.0 to -0.6%						
-0.5 to -0.1%						
0.0-0.4%						
0.5-0.9%						
1.0-1.4%						
1.5-1.9%						
2.0-2.4%						
2.5-2.9%						
3.0-3.4%						
3.5-3.9%						
≥ 4.0%						
Total	100	100	100	100	100	100

* Standardised ESA definition. Probabilities should sum to 100%. Average of the period.

Memorandum item:
In the space below please indicate the main economic factors affecting the outlook for the euro area GDP growth rate over the forecast horizon distinguishing, where possible, between upward and downward revisions each year and main risks associated with the GDP outlook. **In particular, we would be grateful if you could comment on the point in the forecast horizon in which - according to your baseline outlook – the level of real GDP will increase again in a sustained manner (i.e. witness successive quarter-on-quarter growth rates larger than 0.0%).**

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

	Point estimate of euro area unemployment rate* Percentage of labour force					
	2013	**2014**	**2015**	**November 2013**	**November 2014**	**5 years ahead (2017)**
Rate (%)						

* Standardised definition produced by Eurostat.

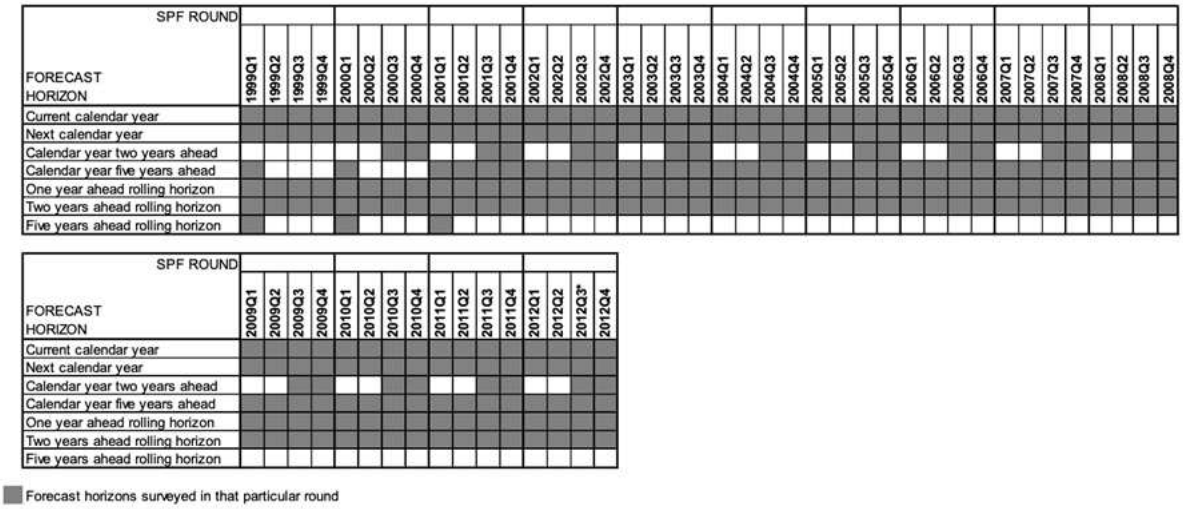
	Probabilities of euro area unemployment rate Percentage of labour force					
	2013	**2014**	**2015**	**November 2013**	**November 2014**	**5 years ahead (2017)**
< 6.5%						
6.5 – 6.9%						
7.0 – 7.4%						
7.5 – 7.9%						
8.0 – 8.4%						
8.5 – 8.9%						
9.0 – 9.4%						
9.5 – 9.9%						
10.0 – 10.4%						
10.5 – 10.9%						
11.0 – 11.4%						
11.5 – 11.9%						
12.0 – 12.4%						
12.5 – 12.9%						
13.0 – 13.4%						
13.5 – 13.9%						
14.0 – 14.4%						
14.5 – 14.9%						
≥ 15.0%						
Total	100	100	100	100	100	100

* Standardised definition produced by Eurostat. Probabilities should sum to 100%. Average of the period.

Memorandum item:
In the space below please indicate the main economic factors affecting the outlook for the euro area unemployment rate over the forecast horizon distinguishing, where possible, between upward and downward revisions each year and main risks associated with the unemployment outlook.

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

Appendix B: Forecast structure over time



Appendix C: Results of full dataset

This part includes the same analysis results as the subsection *re-calibration parameters and forecasting horizon*, with the full data. Tables 11–14 correspond to Tables 4–7 in the main manuscript.

TABLE 11: Inflation by FH (all γs).

FH	n	Mean	Median	SD	IQR
The current year	3402	5.53	1	14.5	2.7
Next year	3315	6.08	0.73	15.86	2.66
Year after next year	1646	8.57	0.91	19.81	5.32
Year 5/6	2008	10.29	1.4	21.93	7.47

TABLE 12: GDP by FH (all γs).

FH	n	Mean	Median	SD	IQR
The current year	3360	3.76	0.84	11.43	2.66
Next year	3289	5.98	1.12	13.19	6.45
Year after next year	1621	5.38	1.01	13.05	4.7
Year 5/6	1987	7.73	1.61	16.42	6.51

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press

TABLE 13: Unemployment by FH (all γs).

FH	n	Mean	Median	SD	IQR
The current year	3207	3.97	0.83	16.11	1.37
Next year	3135	6.87	0.84	23.09	2.77
Year after next year	1518	9.49	1.17	26.28	5.91
Year 5/6	1835	2.52	0	11.75	0.81

TABLE 14: Recalibration parameters for the three indicators for short and long term forecasts.

		n	Mean	Median	SD	IQR
Inflation	Short Term	6716	5.801	0.841	15.187	2.675
	Long Term	3654	9.52	1.174	21.019	6.477
GDP	Short Term	6649	4.856	0.965	12.381	4.393
	Long Term	3608	6.675	1.234	15.041	5.819
Unemployment	Short Term	6342	5.399	0.838	19.92	1.907
	Long Term	3353	5.675	0.334	20	2.19

https://doi.org/10.1017/S1930297500009049 Published online by Cambridge University Press