ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Econometrics & Operations Research

Real-Time Combined Forecasts of the Survey of Professional Forecasters

Date final version: 5 July 2020

Supervisor
P.A. Opschoor

Second assessor
A. Tetereva

Student
Dani S. Klein

ID number
480687

Abstract

In this paper, we elaborate on the research of Diebold and Shin (2019), who use machine learning to combine forecasts. We especially focus on the real-time applicability of their procedures, wherefore we propose transformed methods and show their results in application to point forecasts of the Survey of Professional Forecasters. The data we use contains one-year-ahead forecasts of the real GDP growth, inflation, and unemployment in the Euro-area over the forecasting period 2000Q1-2018Q4. Furthermore, we incorporate forecaster uncertainty in another method, of which we obtain results using density forecasts of the same applications. We find that our real-time applicable point forecast combinations outperform a simple average for inflation. However, the density forecast combinations are unable to significantly beat a simple average.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

The "equal-weights puzzle" represents the phenomenon that combined forecasts, by means of simple-averaging, are often not outperformed by out-of-sample theoretically optimal forecast combinations (Clemen 1989; Diebold 1989). Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2013) show that simple averages are likely to be much better than any individual forecast and to be close to the optimum, even if the simple averages are not fully optimal. However, Diebold and Shin (2019) questioned whether some poor-performing forecasts should not be eliminated before combining them to one comprehensive forecast. Therefore, they propose a LASSO-based procedure and implement quarterly Euro-area real GDP growth point forecasts from the European Central Bank (ECB) Survey of Professional Forecasters (SPF) to test it.

Diebold and Shin (2019) presume that some forecasts could be unwanted due to redundancy, for example. Because of this, they first suggest setting some weights to zero, which they call "selecting to zero". Second, they shrink the remaining weights toward equality. Diebold and Shin (2019) propose the "partially-egalitarian LASSO", which is based on the standard LASSO (Tibshirani 1996) but modified such that the selection and shrinkage properties are included. They find that a good combined forecast is obtained when much forecasters are discarded, and the survivors' forecasts are averaged.

To avoid the obligation to choose an ex post tuning parameter, they suggest subset-averaging procedures that select and average forecasters. These forecast combinations allow for a different number of selected forecasters and number of observations used in the moving window, for each point in time. These two parameters are chosen ex ante and such that the joint forecasting error in the recent past is minimized. Implementing the real GDP growth data into their procedures yields promising results. They show that the optimal forecast combinations on average only select the two best forecasters based on their performances in the two most recent quarters. But a closer look to their research still raises doubts about the real-time applicability of the real GDP application.

We observe that, when evaluating forecasters for selection, Diebold and Shin (2019) use performance measures of forecasts that are unknown ex ante. They are partly based on "future" realizations unknown in real time, causing questionable research conclusions for the specific application. Another flaw regarding real-time applicability is the fact that they select the ex post most frequently responding forecasters. Lastly, final-revised realizations are used for all forecast evaluations, instead of the latest released realizations. To estimate recent (within a year ago) performance measures at a certain point in the evaluation period, one could not have used the final-revised realizations, which sometimes are released only one year after.

Because we only focus on forecasts coming from the SPF, we investigated the further possibilities that come with this survey. A very useful attribute of the surveys is the fact that it does not only contain point forecasts, but also density forecasts. These forecasts could be used to measure the uncertainty that comes with a point forecast of a certain forecaster. Kenny, Kostka, and Masera (2015) examine how SPF density forecast characteristics are linked to its performances. They show that the spread, as well as location, of the density forecasts has a significant impact on their performances. This finding suggests that one could improve the procedures of Diebold and Shin (2019) in a way that uncertainty is considered too when evaluating forecasters.

We contribute to the existing literature by comparing the application of Diebold and Shin (2019) with a real-time applicable application. Here we leave out the last three surveys at each evaluation time, because their corresponding realizations are unknown then. We also do not select the ex post but ex ante most frequently responding forecasters. Furthermore, for recent performance measures at a certain point in the evaluation period, we use realizations from the Real-Time DataBase (Giannone, Henry, Lalik, and Modugno 2012) instead of the final-revised realizations known at the end of this period. Altogether, it should make the application completely applicable in real time, making it usable in practice.

In our research, we also challenge the subset-averaging procedures of Diebold and Shin (2019) with our procedures that consider the uncertainty of forecasters when selecting them. Lastly, we check whether the results of Diebold and Shin (2019) are consistent with the macroeconomic variables inflation and unemployment.

Our research could be of great importance for scientific researchers who focus on combining forecasts using machine-learning procedures, because we discuss and propose many methods in this framework. Furthermore, the ECB may profit from our research, because they will keep a better overview on how to use the SPF to improve macroeconomic forecasts in the Euro-area. Efficient inflation forecasts could also help the ECB target inflation. Lastly, macroeconomists could also use the results of our research, leaving out the need to develop their own predicting models.

We find that our real-time applicable real GDP growth point forecast combinations do not outperform a simple average, which opposes the original results of Diebold and Shin (2019). Although this also holds for the unemployment application, the combinations significantly outperform a simple average when regarding inflation. Nevertheless, the results of the density forecast combinations are less promising because they cannot significantly beat a simple average. However, the methods we use and propose could be fundamental for further research.

This paper proceeds as follows. Section 2 summarizes the methods used by Diebold and Shin (2019), describes the forecast combination procedures based on their results, and introduces our density forecast combination procedure. Section 3 elaborates on the data we implement in all methods, describes the real-time applicability issues, and proposes solutions to these problems. In Section 4, we provide, compare, and discuss the results of the macroeconomic applications. Section 5 concludes, elaborates on the limitations of our research, and proposes recommendations for future research.

2 Methodology

As most of our methods originate from the research of Diebold and Shin (2019), we provide a summary of their approaches. The methods are made such that any balanced panel dataset containing realizations and multiple forecasts of one certain variable can be implemented.

A combined forecast of y_{t+h} is a weighted average of individual forecasts:

$$C_{t+h} = \beta_1 f_{1,t+h} + \beta_2 f_{2,t+h} + \cdots + \left(1 - \sum_{i=1}^{K-1} \beta_i\right) f_{K,t+h}. \tag{1}$$

Bates and Granger (1969) show that minimization of the forecast error, $e_{C_{t+h}} = C_{t+h} - y_{t+h}$, is reached when

$$\beta = \beta^* = \frac{(\Sigma^{-1}\mathbf{i})}{(\mathbf{i}'\Sigma^{-1}\mathbf{i})}, \tag{2}$$

where Σ is the variance-covariance matrix of the individuals' forecast errors $e_{i,t+h} = f_{i,t+h} - y_{t+h}$ and $\mathbf{i}$ is a column vector of ones. We could also obtain β^* from the population regression $y_{t+h} \to f_{1,t+h}, \ldots, f_{K,t+h}$, subject to $\sum_{i=1}^{K} \beta_i = 1$.

The basic framework of the methods discussed in Diebold and Shin (2019) for h-step-ahead forecasts is as follows:

$$\hat{\beta}_m = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_{t+h} - \sum_{i=1}^{K} \beta_i f_{i,t+h} \right)^2 + A_m \right), \tag{3}$$

where penalization A_m depends on which method, m, is used. Standard ridge and LASSO have penalizations $A_{Ridge} = \lambda \sum_{i=1}^{K} \beta_i^2$ and $A_{LASSO} = \lambda \sum_{i=1}^{K} |\beta_i|$, respectively. However, they do not induce the desired properties, because ridge regression shrinks toward zero and does not select at all. LASSO, on the other hand, selects to zero, but shrinks toward zero instead of equality. To shrink the weights toward equality, Diebold and Shin (2019) propose egalitarian ridge and LASSO, whose penalization parts are $A_{eRidge} = \lambda \sum_{i=1}^{K} \left(\beta_i - \frac{1}{K}\right)^2$ and $A_{eRidge} = \lambda \sum_{i=1}^{K} |\beta_i - \frac{1}{K}|$, respectively. Only, in contrast to the standard ridge and LASSO, these methods do not select to zero, while they do shrink toward equality.

To obtain weights that are selected to zero and shrunk toward equality, Diebold and Shin (2019) propose the "partially-egalitarian LASSO", which has penalization

$$A_{peLASSO} = \lambda_1 \sum_{i=1}^{K} |\beta_i| + \lambda_2 \sum_{i=1}^{K} \left(\beta_i - \frac{1}{p(\beta)} \right)^2. \tag{4}$$

In this penalization, $p(\beta)$ represents the number of non-zero β elements. They explain it is hard to implement the peLASSO using one step, because the objective function is discontinuous at $\beta_i = 0$. Therefore, they propose an easier implementable two-step analogue, where, in step 1, k forecasts are selected from the total K forecasts. Subsequently, step 2 shrinks the weights on the k forecasts toward $1/k$. The only method discussed above that can select to zero is LASSO, but two methods obtain the property of shrinking toward equality: eRidge and eLASSO. Hence, we use LASSO in step 1, and we consider eRidge, eLASSO and a simple average for step 2. The difference between eRidge and eLASSO for step 2 is that eLASSO causes the complete procedure to first select some weights to zero, some of the surviving weights to $1/k$ and shrink the rest toward $1/k$.[1]

Just like Diebold and Shin (2019), we use an expanding window for $6 \leq t \leq 20$ and a moving window of 20 periods for $t > 20$. The first five forecasts (for $1 \leq t \leq 5$) are only used to estimate β in next iterations. We use a grid of 200 values for λ: starting with an equally-spaced grid on [-15,15], we obtain a grid on (0, 3269017] by exponentiating the initial grid. We select the optimal tuning parameter λ^* for the different methods ex post, for ease of analysis.

2.1 Forecast combination

In their application to real GDP growth forecasts of the Survey of Professional Forecasters (SPF), Diebold and Shin (2019) show that only a few forecasters are selected in the ex post optimal solution of each peLASSO method. Therefore, they introduce a method that does not need penalty parameter selection, which they call the "average-best forecast combination". The "individual-based average-best N" forecast combination selects the N best performed forecasters at each point in time until then. The forecast combination for one year ahead will then be the simple average of the predictions of these top-performed forecasters.

Due to the ex post selection of N, they also propose an "individual-based average-best $\leq N_{max}$" forecast combination. This method picks the ex-ante optimal N, for $N = 1, \ldots, N_{max}$ at each time, where N_{max} is chosen such that the selection procedure is hardly affected by its boundaries. Although there still is a small ex post effect of the selection of N_{max}, when set large enough, this effect becomes negligibly small.

Using the 20-quarter window, they find that $N = 3$ and $N = 4$ give the lowest RMSEs, matching the ex post optimal peLASSO results, where 2.95 forecasters where selected on average each period. They also find that the "individual-based average-best ≤ 6" forecast combination outperforms a simple average in terms of RMSE, although not significantly. The choice of $N_{max} = 6$ seems fair due to the negligibly small differences with the results of $N_{max} = 5$ and $N_{max} = 4$.

Other combinations Diebold and Shin (2019) propose are the LASSO-based average-best and the best-average combination. But due the slightly worse performance of the former compared to the individual-based approach, we do not consider that combination. We also do not consider the best-average combination because it practically delivers the same results as average-best, and it is computationally heavier.

The last and most real-time applicable forecast combinations they introduce are the best ($\leq$

[1] To measure the performances of all methods, we first check for unbiasedness and variances of their produced forecasts. Second, we combine these measures to the Root Mean Squared Error, which is monotonically increasing in the variance and squared bias (RMSE is the square root of their sum).

N_{max}, W)-average and best ($\leq N_{max}, \leq W_{max}$)-average[2]. These approaches incorporate the fact that choosing a moving window of $W = 20$ quarters is unsubstantiated and might even yield worse results than when using another window width. The results of their best ($\leq 6, W$)-average combinations show an optimum around $W = 2$ and $W = 3$ and the best ($\leq 6, \leq 40$)-average combination also has an average window width of 2.02. Therefore, the initially chosen window width $W = 20$ seems extremely overestimated. The best ($\leq 6, \leq 40$)-average combination turns out to perform better (almost significantly; $p = 0.11$) than a simple average, with an eight percent lower out-of-sample RMSE.

We use the same number of observations for each average-best ($\leq 6, W$) forecast combination for the sake of completeness. Due to scarcity, this does not mean we only start evaluating the forecast combinations after 40 observations (the highest window width considered). It means that we consider the forecasts of the highest possible window for a certain average-best ($\leq 6, W$) combination, when the window width W is bigger than the number of periods that have passed by at a specific time. Although not explicitly mentioned by Diebold and Shin (2019), we think they did this too because it would be consistent with the data implementation of their combining methods. Nevertheless, we realize this procedure slightly decreases the average window width used in the average-best ($\leq 6, \leq 40$) forecast combination.

2.2 Density forecast combination

One might argue that point forecasts do not represent a forecaster's expectations entirely. A seemingly important aspect missing is the uncertainty of his point forecast, measurable using density forecasts. Kenny et al. (2015) show that the spread, as well as location, of density forecasts obtained with the SPF has a significant impact on their performances, suggesting that uncertainty indeed is important. One could also argue this based on intuition. Assume there are two forecasters with the same very low prediction RMSE for a certain variable over the last 20 quarters. One stated at each survey he was very certain his forecasts would come true, but the other forecaster mentioned being less certain with his forecasts. In this case, one might argue that the former forecaster should be given a higher evaluation than the latter when picking the N best forecasters at that time. But also, in case both have a high prediction RMSE, one could argue that the uncertain forecaster should be given a higher evaluation.

Kenny et al. (2015) find that the Ranked Probability Score (RPS) of Epstein (1969) is a fitting measure for SPF density forecast performances. These density forecasts are sets of probabilities assigned to certain ranges of possible outcomes, called "bins". x^j_{t+h} denotes a binary variable which is 1 when the eventual realization at time $t+h$ falls into bin j and 0 otherwise. $X^k_{t+h} = \sum_{j=1}^{k} x^j_{t+h}$ is the cumulative distribution function of x_{t+h}, being 1 when the realization falls into one of the bins $j = 1, \ldots, k$ and 0 otherwise. Likewise, $F^k_{i,t+h} = \sum_{j=1}^{k} \hat{p}^j_{i,t+h}$ is the cumulative distribution function of the estimated bin probability $\hat{p}_{i,t+h}$. The $RPS_{i,t+h}$ then is defined as follows, with K_{max}

[2] We, on the other hand, consider the average-best ($\leq N_{max}, W$) and average-best ($\leq N_{max}, \leq W_{max}$) instead.

representing the number of bins used:

$$RPS_{i,t+h} = \sum_{k=1}^{K_{max}} \left(F_{i,t+h}^{k} - X_{t+h}^{k} \right)^2 . \tag{5}$$

We use the RPSs in the selection parts of average-best N, $\leq N_{max}$, $(\leq N_{max}, W)$ and $(\leq N_{max}, \leq W_{max})$ at each point in time. Instead of picking the N best forecasters based on their RMSE performances, we pick them based on the sum of their previous RPS performances.

3 Data

The data we use comes from the European Central Bank's quarterly Survey of Professional Forecasters[3]. It includes one-year-ahead forecasts of annual real GDP growth, annual inflation, and the level of unemployment over the period 1999Q1-2020Q2. Nevertheless, the forecasts of real GDP growth and the unemployment rate are less than one year ahead, as mentioned by Genre, Kenny, Meyler, and Timmermann (2013). Predictions are made for the periods that begin with the latest available release of each macroeconomic variable. Therefore, "one-year-ahead" forecasts of GDP growth are actually six to eight months ahead, while the unemployment rate forecasts are actually eleven months ahead. However, inflation releases coincide with the launches of the surveys, causing the inflation "one-year-ahead" forecasts to be well-defined.

Some forecasters do not tend to respond to each survey, and while new forecasters enter, other forecasters leave the SPF. This causes an unbalanced dataset, containing many unknown datapoints. To decrease the number of missing values in the data, we filter the set of forecasters to obtain a subset, based on the extent to which they respond to the surveys. As carried out by Diebold and Shin (2019), we select the 23 most frequently responding forecasters over the evaluation period. Although they do not elaborate on the choice of this number, we adopt it to acquire plausible comparisons between their and our results. All empty values within this subset are subsequently estimated as the fit of an AR(1) panel regression:

$$\hat{y}_{i,t+h} - \bar{y}_{t+h} = \beta_{i,t}(\hat{y}_{i,t+h-1} - \bar{y}_{t+h-1}) + \epsilon_{i,t+h}, \tag{6}$$

where β is estimated recursively over the sample period, as proposed by Genre et al. (2013), and $\bar{y}_{t+h}$ is the average prediction for $t+h$ of all forecasters that responded at time t. When a missing prediction, $\hat{y}_{i,t+h}$, occurs at time t of forecaster i, we replace it with $\hat{\beta}_{i,t}(\hat{y}_{i,t+h-1} - \bar{y}_{t+h-1}) + \bar{y}_{t+h}$. We predict what the forecaster would have predicted by using his historical relative optimism or pessimism.

Density forecasts of the SPF, also included in our data, consist of multiple bins of size 0.5% in the middle (closed intervals), and infinitely large bins at the extremities (open intervals). A forecaster should indicate his estimated probabilities of the realization falling into each bin. For

[3] Obtainable at https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/index.en.html

example, the probabilities for an inflation realization y_{t+h} at time $t+h$ that should be predicted at time t are $P(y_{t+h} < -2.0)$, $P(-2.0 \leq y_{t+h} < -1.5)$, …, $P(3.5 \leq y_{t+h} < 4.0)$ and $P(y_{t+h} \geq 4.0)$.

As with the point forecasts, we also have an issue regarding missing values for the density forecasts. To impute these values, we use an approach described by Kenny et al. (2015) that balances the panel based on the above approach for the point forecast imputations (Genre et al. 2013). The method first uses the AR(1) panel regression as in Equation (6) to impute the point forecast, when missing[4]. Using the most recently submitted probabilities of the forecaster, the density forecast is imputed.

In a more complete description, Kenny, Kostka, and Masera (2013) mention some assumptions are made to make the imputations. First, for ease of analysis, they assume that "the probabilities within a given range are uniformly distributed within that range". Furthermore, the open intervals at the edges of the densities are assumed to be closed intervals with an equal width as the closed intervals in the middle, which is 0.5%. This seems like a less valid assumption to make, but when comparing it with other likelier assumptions, they found no notable impact. This coincides with the small amount of times a non-zero probability is assigned to one of the "edge bins" by the forecasters.

Due to the assumptions, one can predict the probabilities a forecaster would have given if he had responded. With the imputed or known point forecast, the increase in forecast compared to the previous (imputed or known) forecast can be calculated: $\Delta\hat{y}_{i,t+h} = \hat{y}_{i,t+h} - \hat{y}_{i,t+h-1}$. For $0 \leq \Delta\hat{y}_{i,t+h} < 0.5$, one can impute "middle bin" probabilities with the following recursion:

$$\hat{p}^{k}_{i,t+h} = \frac{0.5 - \Delta\hat{y}_{i,t+h}}{0.5} \cdot \hat{p}^{k}_{i,t+h-1} + \frac{\Delta\hat{y}_{i,t+h}}{0.5} \cdot \hat{p}^{k-1}_{i,t+h-1}, \tag{7}$$

while for $-0.5 < \Delta\hat{y}_{i,t+h} < 0$ probabilities are moved downwards, so

$$\hat{p}^{k}_{i,t+h} = \frac{0.5 - \Delta\hat{y}_{i,t+h}}{0.5} \cdot \hat{p}^{k}_{i,t+h-1} + \frac{\Delta\hat{y}_{i,t+h}}{0.5} \cdot \hat{p}^{k+1}_{i,t+h-1}. \tag{8}$$

Two other cases are $\Delta\hat{y}_{i,t+h} > 0.5$ and $\Delta\hat{y}_{i,t+h} < -0.5$, for which we use a modulo operation. Let $m = \Delta\hat{y}_{i,t+h} \pmod{0.5}$ be the modulus of the division of $\Delta\hat{y}_{i,t+h}$ by 0.5 and r the remainder, for a specific individual forecast. Then we can impute probabilities for $\Delta\hat{y}_{i,t+h} > 0.5$ using

$$\hat{p}^{k}_{i,t+h} = (1-r) \cdot \hat{p}^{k-m}_{i,t+h-1} + r \cdot \hat{p}^{k-m-1}_{i,t+h-1}, \tag{9}$$

and for $\Delta\hat{y}_{i,t+h} < -0.5$ we use

$$\hat{p}^{k}_{i,t+h} = (1-r) \cdot \hat{p}^{k+m}_{i,t+h-1} + r \cdot \hat{p}^{k+m+1}_{i,t+h-1}. \tag{10}$$

Note that Equation (9) and (10) also hold for $0 \leq \Delta\hat{y}_{i,t+h} < 0.5$ and $-0.5 < \Delta\hat{y}_{i,t+h} < 0$, respec-

[4] A methodological survey of the ECB SPF conducted in 2009 shows that a large majority of the forecasters claims that their point forecasts correspond to the means of their density forecasts (see https://www.ecb.europa.eu/pub/economic-bulletin/mb/html/index.en.html). Therefore, it seems well-grounded to use point forecasts as location indicators of density forecasts.

tively. Therefore, we use Equation (9) for $\Delta\hat{y}_{i,t+h} \geq 0$ and Equation (10) for $\Delta\hat{y}_{i,t+h} < 0$.[5]

Lastly, regarding the data implementation, we only use surveys over the period 1999Q3-2018Q2 for real GDP growth rates and 1999Q2-2018Q1 for inflation and unemployment rates. This causes the forecasting period to be 2000Q1-2018Q4 (exactly 19 years) for all three macroeconomic variables, due to differences in latest releases, as mentioned in Section 3. We choose the latest conducted survey in our data to be 2018Q2 to ensure that the realizations of (especially) the GDP growth rates in the used data vintage are their "final-revised" values, as recommended by Diebold and Shin (2019).[6]

3.1 Real-time applicability issues and solutions

Nevertheless, the results of the forecast combinations of Diebold and Shin (2019) would have been impossible to obtain in practice due to some real-time applicability issues in their application.

First, one should notice that, at each point in time, forecast performances of very recent forecasts are not known in real time. Only the ones that have a forecasting period that ended (and whose realization was released) before the evaluation time could be used. To predict whether an individual forecast will be good, we need to evaluate a forecaster's historical performance. For illustration we focus on the real GDP growth forecasts of 2006Q1, which regard the period 2005Q4-2006Q3 (see Figure 1). These have been made middle 2006M1, so in real time, we only could have used data that was known at 2006M1. The survey of 2005Q1 asked for forecasts of 2004Q4-2005Q3, and the preliminary realization is only known one month after the end of a quarter (Diebold and Shin 2019). This means that the forecasts of survey 2005Q1 only could have been evaluated from the beginning of 2005M11. The forecasts of the subsequent survey, 2005Q2, could not have been evaluated in 2006M1, because even the preliminary realization of the corresponding period, 2005Q1-2005Q4, had not been released at that time. Furthermore, the forecasts of the last two surveys, 2005Q3 and 2005Q4, could not have been evaluated because their corresponding forecasting periods, 2005Q2-

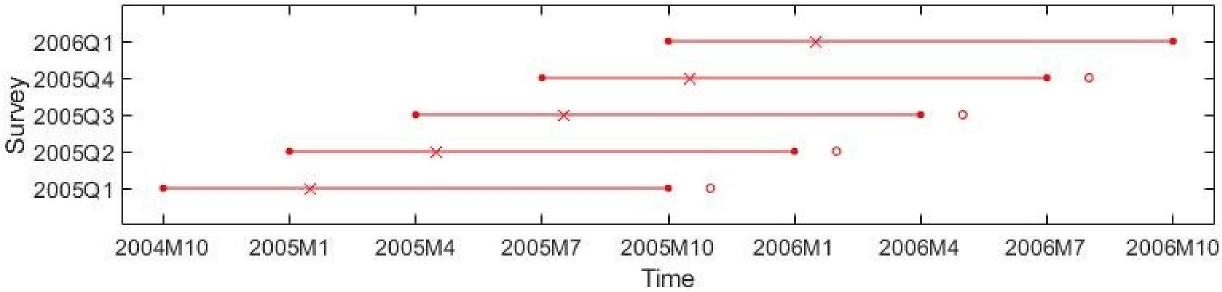


Figure 1: Mapping of the availability of real GDP data over time. For each survey: The red line represents the forecasting period, the moment of response is indicated by the "x" symbol, and the "○" symbol is the release moment of the preliminary realization of the corresponding period.

[5] We also add multiple empty bins before and after the K_{max} bins we have now, such that the imputations can also be made for the "edge" bins and for some of the bins near the extremities when there is a high modulus.

[6] The real GDP data, coming from statistical agency Eurostat, and the data with inflation and unemployment realizations, from the ECB, can be found in the ECB Statistical Data Warehouse at https://sdw.ecb.europa.eu/browse.do?node=9691101.

2006Q1 and 2005Q3-2006Q2, did not even end yet. Thus, "historical" performance measures of the real GDP growth forecasts from the previous three surveys cannot be used in real time.[7]

Diebold and Shin (2019) do not tackle (or address) this issue, wherefore we guess this causes their optimal window width to be very low ($W = 2$ and $W = 3$). To solve this, we leave out the three last surveys when evaluating forecasters each time. Therefore, in the real-time application of the combining methods, we use an expanding window from time 1 until $t - 3$ to estimate β_t for $6 \leq t \leq 23$. For $t > 23$, we use a moving window of 20 periods.

A second inconvenience is the choice of $N_{max} = 6$, because this is based on the results of the best $(N, 20)$-average combinations. However, it might be inconsistent with the results of the best $(N, 2)$-average forecast combinations. Therefore, we assess possible structural differences between multiple window widths and base the range of allowed numbers of selected forecasters on that.

The third issue regarding real-time applicability, is the fact that the 23 forecasters are picked based on the number of times all forecasters have responded to the surveys over the full sample. Unfortunately, implementing this procedure in real time is impossible, since one could not have known which forecasters shall respond most often in future surveys. As a solution to this problem, we use the same windows as with the forecast evaluations (expanding window for $6 \leq t \leq 20$ and moving window of 20 observations for $t > 20$) to pick the 23 most responding forecasters in the current past. It causes the panel to still be unbalanced, but the used data to be balanced, because we do not have any missing values in the final data we implement at each point in time. This approach also possibly lowers the number of forecasts that need to be imputed, because forecasters in general seem to have periods of many and little responses separated over time.

The last inconvenience in the application is the used estimates of the macroeconomic variables. As Diebold and Shin (2019) already point out themselves, "additional non-standard revisions sometimes occur after more than 100 days". They also mention they wait approximately a year such that their data set almost only contains final-revised estimations. Therefore, at each point in time, we consider estimates of the Real-Time DataBase of Giannone et al. (2012) for the last four quarters in our real-time applications. These realizations are released two months after the end of a forecasting period. For all previous quarters we use the approximately "final-revised" values.

4 Results

In this section, we compare results of the real GDP forecasts based on our real-time applicable methods with the results based on the original methods of Diebold and Shin (2019). We also do this for the main results of inflation and unemployment forecasts[8] and we discuss similarities and differences between the conclusions regarding the three macroeconomic variables. Finally, we investigate whether our density forecast combinations can improve a simple average or the average-best $(\leq N_{max}, \leq W_{max})$ forecast combination.

[7] Likewise, this holds for the unemployment and inflation forecasts.

[8] See Appendix A for all results of the real GDP, inflation and unemployment applications.

4.1 Ex post combining methods

4.1.1 Original

Table 1 shows performance measures of all combining methods, some individual quantiles, and a simple average for the real GDP growth. The results are like the results of Diebold and Shin (2019), but unequal due to the slightly different evaluation period and possible realization changes in late revisions. However, many of their conclusions still hold: No method outperforms the ex post best individual forecaster regarding accuracy (RMSE = 1.35) and, besides LASSO, unbiasedness (Bias = 0.02). eRidge and eLASSO approximately produce a simple average due to high optimal shrinkage strength, wherefore their performances coincide. This heavy regularization also explains the equal performances of the three peLASSO methods.[9] Just as with their results, the peLASSO methods do not significantly outperform a simple average, when guided by the DM statistics.[10]

Table 1: Combining method performance measures for real GDP based on ex post optimal penalty parameters λ^*.

Regularization Group	RMSE	Bias	Variance	λ^*	#[1]	DM[2]	p-val[3]
Ridge	1.44	-0.11	2.08	40.18	23.00	-0.43	0.66
LASSO	1.46	0.00	2.15	0.38	2.90	-0.48	0.68
eRidge	1.41	0.05	2.00	max	23.00	$-\infty$	1.00
eLASSO	1.40	0.02	1.99	2.66	23.00	1.38	0.08
peLASSO (Average)	1.38	0.07	1.93	0.15	3.93	0.51	0.31
peLASSO (eRidge)	1.38	0.05	1.93	(0.15, max)	3.93	0.51	0.31
peLASSO (eLASSO)	1.38	0.06	1.93	(0.15, 3.10)	3.93	0.52	0.30
Comparisons	RMSE	Bias	Variance	λ^*	#	DM	p-val
Best	1.35	0.02	1.86	N/A	1	0.78	0.22
90%	1.42	-0.07	2.03	N/A	1	-0.43	0.67
Median	1.43	0.12	2.06	N/A	1	-0.31	0.62
10%	1.50	-0.07	2.28	N/A	1	-1.38	0.92
Worst	1.64	-0.08	2.71	N/A	1	-1.53	0.94
Average	1.41	0.05	2.00	N/A	23	N/A	N/A

[1] # is the average number of forecasters selected;
[2] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey, Leybourne, and Newbold (1997); [3] p-val represents the p-value.

Nevertheless, there are some differences in the results. Ridge and LASSO seem to perform worse than a simple average, whereas Diebold and Shin (2019) show that they perform approximately equally well, despite their shrinkage toward zero. Furthermore, the RMSE of the peLASSO methods is not seven percent lower (Diebold and Shin 2019), but only two percent. This is caused by the lower variance of the forecast errors at the small expense of a slightly higher bias. Another dissimilarity is the average optimal number of forecasters selected when using the peLASSO methods, which is # = 3.93 in our results, much higher than in the results of Diebold and Shin (2019) (# = 2.95).

[9] Each peLASSO method is from now on referred to with the method used in its second step in parentheses.

[10] However, one should bear in mind that almost identical forecasts cause unreliable DM statistics, because the loss differential $e^2_{i,t} - e^2_{j,t}$ then has a variance and mean of approximately zero. The statistic is also unreliable when many or strong outliers occur, which violate the DM assumption of normally distributed squared forecast errors.

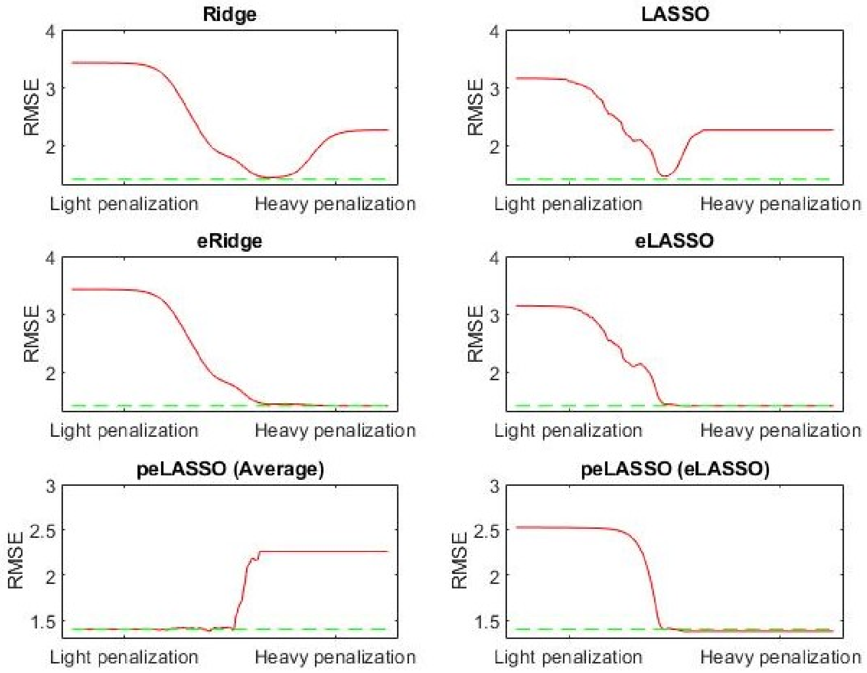


Figure 2: RMSEs as function of λ for various forecast combination methods for real GDP, with the dashed line representing a simple average.

One could also see that we obtain slightly different but comparable ex post optimal λ^*s compared to Diebold and Shin (2019). Figure 2 shows that the RMSEs as functions of λ also are approximately equal to the functions of Diebold and Shin (2019) for all methods.

Table 2: peLASSO (Average) and simple average forecast performance measures based on ex post optimal penalty parameters λ^*.

Variable	Method	RMSE	Bias	Variance	λ^*	#[1]	DM[2]	p-val[3]
Real GDP	peLASSO	1.38	0.07	1.93	0.15	3.93	0.51	0.31
	Average	1.41	0.05	2.00	N/A	23	N/A	N/A
Inflation	peLASSO	0.83	0.08	0.69	0.32	2.21	2.19	0.01
	Average	0.87	0.06	0.77	N/A	23	N/A	N/A
Unemployment	peLASSO	0.63	-0.07	0.40	1.97	2.51	2.02	0.02
	Average	0.68	0.08	0.46	N/A	23	N/A	N/A

[1] # is the average number of forecasters selected; [2] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997);
[3] p-val represents the p-value.

In Table 2 we see that the results of the peLASSO (Average) method for inflation and unemployment forecasts coincide with those for the real GDP growth rate. The peLASSO forecasts outperform simple averages even on a significance level of 0.05. Furthermore, considering the transition from simple-averaging to the peLASSO (Average), the RMSEs even decrease by five and seven percent for inflation and unemployment, respectively. Again, this is caused by the lower variances of the forecast errors. Last, we notice that all macroeconomic variable values are overestimated on average, given the small positive bias in the simple-average predictions. But because of the linear quadratic dependence of the RMSEs on the biases and the fact that all biases have a magnitude below 0.10, the RMSEs are mostly driven by the forecast variances.

4.1.2 Real-time applicable

Table 3, Table 4 and Figure 3 show the results of the real-time applications described in Section 3.1. The first observations we make when comparing Table 3 with Table 1 are the higher bias (0.11), variance (1.98), and therefore RMSE (1.40) of the peLASSO methods. Therefore, solving the real-time applicability issues seems to worsen the performance of these methods. We suppose this is mainly due to omitting the last three quarters when evaluating the forecasters. However, the impact is not that big due to the large window width $W = 20$ used for the methods, causing the forecasters to be hardly evaluated on their last three forecasts.

Table 3: Real-time applicable combining method performance measures for real GDP based on ex post optimal penalty parameters λ^*.

Regularization Group	RMSE	Bias	Variance	λ^*	#[1]	DM[2]	*p*-val[3]
Ridge	1.51	-0.03	2.31	115.44	23.00	-4.98	1.00
LASSO	1.72	0.15	2.97	0.59	2.29	-1.48	0.93
eRidge	1.41	0.03	2.02	max	23.00	-13.18	1.00
eLASSO	1.41	0.03	2.01	3.60	23.00	3.19	0.00
peLASSO (Average)	1.40	0.11	1.98	0.11	4.60	0.39	0.35
peLASSO (eRidge)	1.40	0.11	1.98	(0.11, max)	4.60	0.39	0.35
peLASSO (eLASSO)	1.40	0.11	1.98	(0.11, 4.19)	4.60	0.39	0.35
Comparisons	RMSE	Bias	Variance	λ^*	#	DM	*p*-val
Best	1.36	0.01	1.87	N/A	1	0.78	0.22
90%	1.41	0.02	2.00	N/A	1	0.54	0.29
Median	1.45	0.11	2.12	N/A	1	-0.63	0.74
10%	1.54	-0.01	2.40	N/A	1	-0.83	0.80
Worst	1.63	-0.09	2.70	N/A	1	-1.45	0.93
Average	1.41	0.03	2.01	N/A	23	N/A	N/A

[1] # is the average number of forecasters selected; [2] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997);
[3] *p*-val represents the *p*-value.

Even when using only real-time applicable data, we observe similar results as with the original application. Ridge and LASSO still perform worse than a simple average, while eRidge and eLASSO approximately produce a simple average. Furthermore, the peLASSO methods coincide with each other, but still do not significantly improve a simple average according to the DM statistics.

A minor difference is that the peLASSO methods select slightly more forecasters on average ($\# = 4.60$) than in the original application ($\# = 3.93$), caused by the slightly lower optimal penalty parameter ($\lambda^* = 0.15$ instead of 0.11). A lower penalty parameter means a less powerful selection to zero in the first step (LASSO), and thus more selected forecasters. This might be due to the real-time differences, but it could also be coincidental because of minor changes around the optimum.

Figure 3 shows the RMSEs as a function of λ, as in Figure 2, but for the real-time application of the real GDP growth. We observe that the functions for Ridge and LASSO practically moved

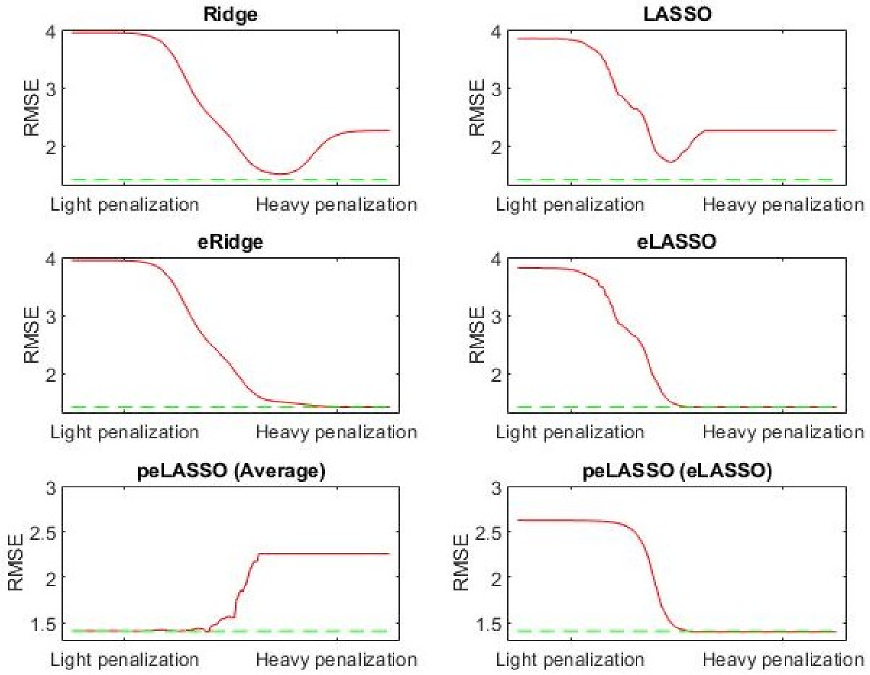


Figure 3: RMSEs as function of λ for various forecast combination methods for real GDP, with the dashed line representing a simple average (real-time applicable).

upwards, although the shape is somewhat retained. Furthermore, the functions of eRidge and eLASSO hardly changed, except for the "starting RMSEs" occurring at light penalizations. Lastly, the peLASSO (eLASSO) function does not seem to have changed, while the peLASSO (Average) shows some deterioration. There are less ex post λ's for which the method outperforms the simple average, suggesting that the chance that this outperforming is coincidental is higher.

Table 4: Real-time applicable peLASSO (Average) and simple average forecast performance measures based on ex post optimal penalty parameters λ^*.

Variable	Method	RMSE	Bias	Variance	λ^*	#[1]	DM[2]	p-val[3]
Real GDP	peLASSO	1.40	0.11	1.98	0.11	4.60	0.39	0.35
	Average	1.41	0.03	2.01	N/A	23	N/A	N/A
Inflation	peLASSO	0.84	0.08	0.71	0.02	7.71	10.65	0.00
	Average	0.86	0.06	0.75	N/A	23	N/A	N/A
Unemployment	peLASSO	0.67	0.07	0.45	0.01	14.94	10.41	0.00
	Average	0.68	0.08	0.46	N/A	23	N/A	N/A

[1] # is the average number of forecasters selected;
[2] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [3] p-val represents the p-value.

Table 4 shows that not only the real-time peLASSO applications of the real GDP contain higher RMSEs than the original results, but also the inflation and unemployment applications. Another difference between Table 4 and Table 2 are the optimal penalty parameters for inflation and unemployment. Where we originally had $\lambda^* = 0.32$ and 1.97, we now observe $\lambda^* = 0.02$ and

0.01, respectively for inflation and unemployment. The decrease of λ^* is much stronger for these applications than for real GDP. This causes the average optimal selected number of inflation and unemployment forecasters to be $\# = 7.71$ and 14.94 instead of $\# = 2.21$ and 2.51, respectively. It indicates that less forecasters should be discarded than originally suspected.

4.2 Average-best forecast combination

4.2.1 Original

Table 5 shows the results of the individual-based average-best N and $\leq N_{max}$ forecast combination procedures mentioned in Section 2.1. They correspond to the results of Diebold and Shin (2019) in a way that an asymptote of the RMSEs is reached at approximately $N_{max} = 4$. Also, an optimal RMSE occurs around $N = 2, 3$ and 4. However, there are differences between the average number of forecasters used in the "individual-based average-best $\leq N_{max}$"; many more forecasters are selected for each N_{max} in our results, compared to the results of Diebold and Shin (2019). However, this corresponds to the results obtained using the peLASSO methods in Table 1. Another remarkable dissimilarity is the fact that the RMSEs increase in N_{max}, meaning that the optimal ex post N_{max} would be 1. However, we see that a higher N_{max} is associated with a lower forecast bias but a higher variance. Therefore, one should make a trade-off between bias and variance to choose the better ex post forecast combination.[11]

Table 5: Average-best forecast combination for real GDP.

Average-best N	RMSE	Bias	Variance	#[1]	DM[2]	p-val[3]
$N = 1$	1.38	0.13	1.90	1.00	0.25	0.40
$N = 2$	1.35	0.11	1.81	2.00	1.62	0.05
$N = 3$	1.37	0.11	1.88	3.00	0.64	0.26
$N = 4$	1.36	0.11	1.83	4.00	1.42	0.08
$N = 5$	1.37	0.12	1.86	5.00	1.17	0.12
$N = 6$	1.37	0.11	1.87	6.00	1.30	0.10
Average-best $\leq N_{max}$	RMSE	Bias	Variance	#	DM	p-val
$N_{max} = 1$	1.38	0.13	1.90	1.00	0.25	0.40
$N_{max} = 2$	1.39	0.10	1.91	1.99	0.45	0.33
$N_{max} = 3$	1.40	0.10	1.96	2.44	0.01	0.49
$N_{max} = 4$	1.41	0.10	1.97	3.13	0.00	0.50
$N_{max} = 5$	1.41	0.09	1.99	3.46	-0.17	0.57
$N_{max} = 6$	1.41	0.09	1.99	3.46	-0.17	0.57
Average	1.41	0.05	2.00	23	N/A	N/A

[1] # is the average number of forecasters selected;
[2] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [3] p-val represents the p-value.

Table 6 shows the results of forecast combination methods average-best $(\leq 6, W)$ and $(\leq 6, \leq 40)$. It points out that a very small window width, optimally $W = 2$, is preferred. Also, the average window width of the average-best $(\leq 6, \leq 40)$ is $W = 2.27$, and the average selected number of

[11] Nevertheless, we will consider $N_{max} = 6$ in our forecast combination to yet reduce the ex post aspect and allow for comparisons with the results of Diebold and Shin (2019).

forecasters is $N = 2.40$. These values are approximately equal to the values from the results of Diebold and Shin (2019), which were $W = 2.02$ and $N = 1.38$. As presumed earlier in Section 3.1, this low average window width is probably caused by the real-time issue regarding the use of the last three quarters. The DM-statistic and its p-value slightly differ with their results; the average-best $(\leq 6, \leq 40)$ now outperforms a simple-average, even on a significance level of 0.05. However, the decrease in RMSE with respect to a simple average is in both results approximately eight percent.

Table 6: Forecast combination for real GDP.

Average-best $(\leq 6, W)$	RMSE	Bias	Variance	#N^{1}	#W^{2}	DM^{3}	p-val^{4}
$W = 1$	1.30	0.07	1.70	2.64	1	1.11	0.13
$W = 2$	1.29	0.07	1.65	2.66	2	1.86	0.03
$W = 3$	1.30	0.11	1.67	1.93	3	1.62	0.05
$W = 5$	1.30	0.09	1.68	3.26	5	3.07	0.00
$W = 7$	1.34	0.09	1.79	2.66	7	0.92	0.18
$W = 10$	1.38	0.08	1.89	3.63	10	0.51	0.30
$W = 15$	1.40	0.10	1.94	3.37	15	0.13	0.45
$W = 20$	1.41	0.09	1.97	3.66	20	-0.03	0.51
$W = 30$	1.42	0.09	2.00	2.46	30	-0.14	0.56
$W = 40$	1.45	0.13	2.09	2.53	40	-0.39	0.65
Average-best $(\leq 6, \leq 40)$	1.29	0.08	1.68	2.40	2.27	1.78	0.04
Average	1.41	0.05	2.00	23	N/A	N/A	N/A

[1] $\#N$ is the average number of forecasters selected; [2] $\#W$ is the average window width selected; [3] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [4] p-val represents the p-value.

Table 7: Forecast combination.

Variable	Method	RMSE	Bias	Variance	#N^{1}	#W^{2}	DM^{3}	p-val^{4}
Real GDP	Average-best $(\leq 6, \leq 40)$	1.29	0.08	1.68	2.40	2.27	1.78	0.04
	Average	1.41	0.05	2.00	23	N/A	N/A	N/A
Inflation	Average-best $(\leq 6, \leq 40)$	0.81	0.07	0.66	2.07	8.76	8.07	0.00
	Average	0.87	0.06	0.77	23	N/A	N/A	N/A
Unemployment	Average-best $(\leq 6, \leq 40)$	0.58	0.03	0.35	1.61	4.73	4.79	0.00
	Average	0.68	0.08	0.46	23	N/A	N/A	N/A

[1] $\#N$ is the average number of forecasters selected; [2] $\#W$ is the average window width selected; [3] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [4] p-val represents the p-value.

Table 7 presents the average-best $(\leq 6, \leq 40)$ forecast combination performances of the inflation and unemployment application too. When comparing them with the results of the real GDP application, we first see that the decrease in RMSEs of the combinations relative to a simple average are consistent. For the inflation forecasts, this decrease is seven percent, while for the unemployment forecasts no less than 14 percent. Another remarkable observation we make is the much higher average window width in both applications, especially for inflation. This is promising for the real-time applications, because the effect of the last three quarters is much smaller and earlier quarters do seem to influence the optimal selection of forecasters. Furthermore, very few forecasters are selected for the latter two applications on average, even less than for the real GDP.

4.2.2 Real-time applicable

One can see in Table 8 that the results of the average-best N combinations are much worse than in the original real GDP application (see Table 5). Not even a single forecast combination outperforms a simple average. Because the minimum RMSE is reached for $N = 23$ (a simple average), which caused no asymptote in the average-best $\leq N_{max}$ RMSEs as function of N_{max}, we allow for all N's ($N_{max} = 23$). In this so-called average-best *all* combination, the average selected number of forecasters is no less than $\# = 10.67$, much higher than in the original application ($\# = 3.46$). We also see that its RMSE is higher than that of a simple average, due to a higher bias *and* variance.

Table 8: Real-time applicable average-best forecast combination for real GDP.

Average-best N	RMSE	Bias	Variance	#[1]	DM[2]	p-val[3]
$N = 1$	1.57	0.16	2.39	1.00	-0.78	0.78
$N = 3$	1.47	0.13	2.09	3.00	-1.11	0.87
$N = 5$	1.46	0.10	2.05	5.00	-2.65	1.00
$N = 7$	1.45	0.09	2.04	7.00	-4.97	1.00
$N = 10$	1.45	0.08	2.03	10.00	-5.93	1.00
$N = 15$	1.43	0.05	1.99	15.00	-2.76	1.00
$N = 20$	1.43	0.03	1.99	20.00	-1.79	0.96
Average-best *all*	1.48	0.09	2.19	10.67	-1.33	0.91
Average	1.41	0.03	2.01	23	N/A	N/A

[1] # is the average number of forecasters selected; [2] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [3] p-val represents the p-value.

Table 9: Real-time applicable forecast combination for real GDP.

Average-best (*all*, W)	RMSE	Bias	Variance	#N[1]	#W[2]	DM[3]	p-val[4]
$W = 1$	1.45	0.05	2.10	14.13	1	-3.18	1.00
$W = 2$	1.42	0.05	2.02	14.21	2	-1.12	0.87
$W = 3$	1.46	0.05	2.13	8.59	3	-2.32	0.99
$W = 5$	1.43	0.07	2.05	4.43	5	-1.87	0.97
$W = 7$	1.44	0.10	2.07	13.96	7	-0.77	0.78
$W = 10$	1.42	0.07	2.02	9.41	10	-0.35	0.64
$W = 15$	1.44	0.09	2.06	8.77	15	-0.58	0.72
$W = 20$	1.50	0.11	2.24	10.37	20	-0.68	0.75
$W = 30$	1.47	0.14	2.15	12.03	30	-0.39	0.65
$W = 40$	1.47	0.14	2.16	12.03	40	-0.40	0.65
Average-best (*all*, ≤ 40)	1.47	0.08	2.18	2.43	6.64	-0.60	0.73
Average	1.41	0.03	2.01	23	N/A	N/A	N/A

[1] $\#N$ is the average number of forecasters selected; [2] $\#W$ is the average window width selected; [3] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [4] p-val represents the p-value.

Table 9 shows that allowing for different window widths in the average-best (*all*, W) and (*all*, ≤ 40) still does not lead to an improvement relative to a simple average. Furthermore, although approximately the same number of forecasters is selected as in the original application, the average window width is somewhat bigger. So instead of selecting forecasters unjustly based on their performances in the last 2 quarters, they are selected based on the third last to $3 + 6.64 \approx$ tenth last quarters. We also observe that even the ex post optimal window widths $W = 2$ or $W = 10$ are not able to perform optimally. Therefore, we can conclude that the real-time real GDP application of the average-best (all, ≤ 40) forecast combination does not outperform a simple average.

Table 10: Real-time applicable forecast combination.

Variable	Method	RMSE	Bias	Variance	#N[1]	#W[2]	DM[3]	p-val[4]
Real GDP	Average-best (all, ≤ 40)	1.47	0.08	2.18	2.43	6.64	-0.60	0.73
	Average	1.41	0.03	2.01	23	N/A	N/A	N/A
Inflation	Average-best (all, ≤ 40)	0.83	0.07	0.69	1.54	6.50	2.19	0.01
	Average	0.86	0.06	0.75	23	N/A	N/A	N/A
Unemployment	Average-best (all, ≤ 40)	0.69	0.00	0.48	4.79	5.90	-0.53	0.70
	Average	0.68	0.08	0.46	23	N/A	N/A	N/A

[1] $\#N$ is the average number of forecasters selected; [2] $\#W$ is the average window width selected;
[3] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [4] p-val represents the p-value.

Table 10 shows that the results of the real GDP application are not consistent with those of inflation. As expected for inflation, (see Section 4.2.1) we see that the impact of the real-time issues is very small regarding the RMSE. It even seems that the average-best (all, ≤ 40) forecast combination significantly outperforms a simple average, although just a three percent lower RMSE is obtained. It is noticeable that the combination on average only selects $\#N = 1.54$ forecasters, while the average window width is approximately the same as for real GDP. Although the latter also holds for unemployment, a lot more forecasters are selected in this application ($\#N = 4.79$). However, the forecast combination does not improve a simple average, just like with the real GDP. We conclude that the average-best (all, ≤ 40) forecast combination only outperforms a simple average in the inflation application.

4.3 Average-best density forecast combination

Table 11 presents the results of the average-best ($\leq 6, \leq 40$), density average-best ($\leq 6, \leq 40$) forecast combination, and a simple average, using the original procedures. The two methods have an approximately equal average number of selected forecasters and window width. However, according to the RMSEs, a simple average is outperformed by both, while the average-best ($\leq 6, \leq 40$) combination seems best for all variables. It even performs significantly better than a simple average on a significance level of 0.05 in all three cases. Although the density average-best ($\leq 6, \leq 40$) has lower RMSEs than a simple average, all RMSEs are higher than those of the average-best ($\leq 6, \leq 40$), and the significance according to the DM statistics is lower.

Table 11: Density and point forecast combination compared with a simple average.

Variable	Method	RMSE	Bias	Variance	#N[1]	#W[2]	DM[3]	p-val[4]
Real GDP	Average-best ($\leq 6, \leq 40$)	1.29	0.08	1.68	2.40	2.27	1.78	0.04
	Density average-best ($\leq 6, \leq 40$)	1.38	0.06	1.92	1.64	2.34	0.27	0.39
	Average	1.41	0.00	2.03	23	N/A	N/A	N/A
Inflation	Average-best ($\leq 6, \leq 40$)	0.81	0.07	0.66	2.07	8.76	8.07	0.00
	Density average-best ($\leq 6, \leq 40$)	0.83	0.09	0.69	2.19	8.13	1.34	0.09
	Average	0.87	0.05	0.76	23	N/A	N/A	N/A
Unemployment	Average-best ($\leq 6, \leq 40$)	0.58	0.03	0.35	1.61	4.73	4.79	0.00
	Density average-best ($\leq 6, \leq 40$)	0.64	0.03	0.41	1.90	3.64	4.68	0.00
	Average	0.70	0.09	0.48	23	N/A	N/A	N/A

[1] $\#N$ is the average number of forecasters selected; [2] $\#W$ is the average window width selected;
[3] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [4] p-val represents the p-value.

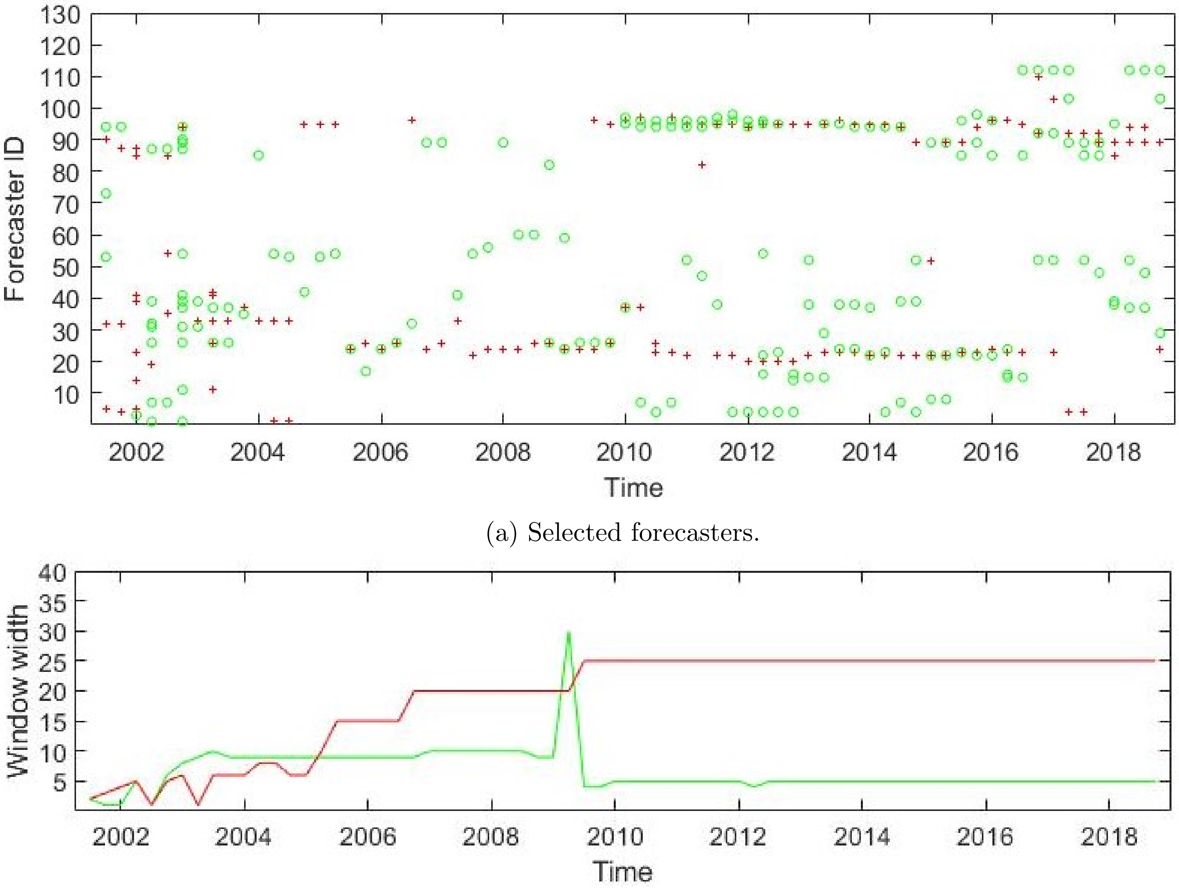
The results of the real-time applications of the three forecast combinations are shown in Table 12. Many differences with the original results of Table 11 occur: As earlier concluded in Section 4.2.2, only the inflation could be better predicted using the average-best (*all*, $\leq$ 40) forecast combination instead of a simple average. One might notice the different simple average RMSE compared to Table 10, but this is due to the fact we now use a set of forecasters that responded most often to the density forecasts each time, not the point forecasts. Therefore, the RMSE of a simple average for unemployment now is higher than the average-best (*all*, $\leq$ 40) combination RMSE. Nevertheless, this minor improvement is not significant according to the DM statistic.

Table 12: Real-time applicable density and point forecast combination compared with a simple average.

Variable	Method	RMSE	Bias	Variance	#N[1]	#W[2]	DM[3]	p-val[4]
Real GDP	Average-best (*all*, $\leq$ 40)	1.47	0.08	2.18	2.43	6.64	-0.60	0.73
	Density average-best (*all*, $\leq$ 40)	1.40	0.18	1.95	1.80	18.97	0.17	0.43
	Average	1.41	0.00	2.03	23	N/A	N/A	N/A
Inflation	Average-best (*all*, $\leq$ 40)	0.83	0.07	0.69	1.54	6.50	2.19	0.01
	Density average-best (*all*, $\leq$ 40)	0.85	0.06	0.73	2.01	11.61	0.48	0.32
	Average	0.87	0.05	0.76	23	N/A	N/A	N/A
Unemployment	Average-best (*all*, $\leq$ 40)	0.69	0.00	0.48	4.79	5.90	-0.53	0.70
	Density average-best (*all*, $\leq$ 40)	0.69	0.01	0.48	3.04	14.83	0.60	0.27
	Average	0.70	0.09	0.48	23	N/A	N/A	N/A

[1] $\#N$ is the average number of forecasters selected; [2] $\#W$ is the average window width selected;
[3] DM is the one-sided statistic against a simple average (Diebold and Mariano 2002), computed as per Harvey et al. (1997); [4] p-val represents the p-value.

For each variable, we also observe a very high window width for the density average-best (*all*, $\leq$ 40) forecast combinations. No less than $\#W = 18.97$ observations are on average used to select the best real GDP forecasters. However, the number of selected forecasters only slightly increases relative to the original results (now between $\#N = 1.80$ and 3.04). Although the density average-best (*all*, $\leq$ 40) has the lowest RMSE for the real GDP application, it is not a significantly better than combination than just a simple average. This also holds for both other applications. The only significantly improving combination is the average-best (*all*, $\leq$ 40) inflation forecast combination.



(a) Selected forecasters.

(b) Window width.

Figure 4: Characteristics of the density average-best (*all*, $\leq$ 40) forecast combinations (red and "+") and the average-best (*all*, $\leq$ 40) forecast combinations (green and "∘") for real GDP.

Figure 4a shows how the set of selected forecasters changes over time in the (density) average-best (*all*, $\leq$ 40) forecast combinations for real GDP. We observe many differences between the two evolving sets, even though the average number of selected forecasters is approximately two for both procedures. We also see for both procedures that relatively a lot of forecasters are selected in the first years, indicating that too little is known about each forecaster to obtain stable evaluations. A combination comprised of many forecasters is more desirable because it gives more certainty.

In Figure 4b we see that the number of observations used to evaluate forecasters in the density average-best (*all*, $\leq$ 40) forecast combination is increasing over time. This is due to the absence of high window width performances in the first few years. The big window used in later periods indicates that density forecast combination performances of certain forecasters do not change rapidly over time. However, the average-best (*all*, $\leq$ 40) only uses five observations after 2010. The large window that follows the Great Recession (2007-2009) looked optimal probably because the impact of high forecast errors around that period on performance measures is smallest for large windows.

Combining both figures also yields some interesting observations. One could see that there is an asymptote in the functions, and that the numbers of selected forecasters for both procedures do not change much after the first few years. This is caused by the use of an expanding window to evaluate the performances of all W's and N's. Thus, the optimal W and N at the latest periods actually indicate the optimal W and N over the total evaluation period. Therefore, the optimal number of observations one should use to combine forecasts depends on the procedure. However, for both forecast combinations, 2-3 forecasters are selected optimally.

One might wonder why the average number of selected forecasters is so small in the procedures, while a simple average can barely be outperformed significantly. One could expect this average to yield very different results, because it selects much more forecasters (all 23). This seemingly contradictory result might therefore be due the "equal-weights puzzle". However, we have shown that a simple average can be outperformed, although not always significantly. Further research could be done on finding a procedure that always significantly outperforms a simple average.

5 Conclusion

In this paper, we elaborate on the research of Diebold and Shin (2019), who use machine learning to combine forecasts. We propose real-time applicable procedures and show their results in application to point forecasts of the Survey of Professional Forecasters. We focus on one-year-ahead forecasts of the real GDP growth, inflation, and unemployment in the Euro-area over the forecasting period 2000Q1-2018Q4. Furthermore, we incorporate forecaster uncertainty in another method, of which we obtain results using density forecasts of the same applications.

We find that our real-time applicable real GDP growth point forecast combinations do not outperform a simple average, which opposes the original results of Diebold and Shin (2019). Although this also holds for the unemployment application, the combinations significantly outperform a simple average when applied to inflation. Nevertheless, the results of the density forecast combinations are less promising, because they are not able to significantly beat a simple average. Still, the methods we propose could be a stepping stone for further research into the topic of forecast combinations.

Although most real-time applicability issues are resolved within our methods, there are still some limitations to our research. First, we do not examine the effect of recessions on our conclusions. Because they generally involve higher forecasting errors and because our performance measures could be sensitive to outliers, these measures do not adequately describe performance over time. Also, we only focus on very specific applications of our methods, while it would be interesting to know whether the conclusions are robust over all data characteristics.

Next to considering these limitations, further research could be done on developing a new method that combines point and density forecast procedures. This could improve performances because the point forecast procedures leave out forecaster uncertainty, while the density forecast procedures might focus too much on it. Some forecasters could be modest about the certainty of their point forecasts, causing underestimation of their ability to forecast. To solve this problem, one could also try to construct a density forecast method that compares forecaster uncertainty with his previous uncertainties (time-series), not with other forecasters' uncertainties (cross-sectional).

References

Aruoba, S. B., Diebold, F. X., Nalewaik, J., Schorfheide, F., & Song, D. (2013). Improving us gdp measurement: A forecast combination perspective, In *Recent advances and future directions in causality, prediction, and specification analysis*. Springer.

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

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

Diebold, F. X. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, *5*(4), 589–592.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, *35*(4), 1679–1691.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*(6), 985–987.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121.

Giannone, D., Henry, J., Lalik, M., & Modugno, M. (2012). An area-wide real-time database for the euro area. *Review of Economics and Statistics*, *94*(4), 1000–1013.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291.

Kenny, G., Kostka, T., & Masera, F. (2013). Can macroeconomists forecast risk? event-based evidence from the euro area spf.

Kenny, G., Kostka, T., & Masera, F. (2015). Density characteristics and density forecast performance: A panel analysis. *Empirical Economics*, *48*(3), 1203–1231.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

A Results

A.1 Ex post combining methods

	Real GDP							*Inflation*						
Regularization Group	RMSE	Bias	Var.	λ^*	#	DM	p-val	RMSE	Bias	Var.	λ^*	#	DM	p-val
Ridge	1.44	-0.11	2.08	40.18	23.00	-0.43	0.66	0.86	-0.07	0.75	1.97	23.00	0.16	0.44
LASSO	1.46	0.00	2.15	0.38	2.90	-0.48	0.68	0.84	0.02	0.72	0.06	4.77	0.50	0.31
eRidge	1.41	0.05	2.00	max	23.00	$-\infty$	1.00	0.86	-0.08	0.75	1.97	23.00	0.15	0.44
eLASSO	1.40	0.02	1.99	2.66	23.00	1.38	0.08	0.85	-0.06	0.74	0.06	23.00	0.29	0.38
peLASSO (Average)	1.38	0.07	1.93	0.15	3.93	0.51	0.31	0.83	0.08	0.69	0.32	2.21	2.19	0.01
peLASSO (eRidge)	1.38	0.05	1.93	(0.15, max)	3.93	0.51	0.31	0.82	0.04	0.68	(0.32, max)	2.21	2.52	0.01
peLASSO (eLASSO)	1.38	0.06	1.93	(0.15, 3.10)	3.93	0.52	0.30	0.83	0.08	0.69	(0.32, 2.29)	2.21	2.19	0.01
Comparisons	RMSE	Bias	Var.	λ^*	#	DM	p-val	RMSE	Bias	Var.	λ^*	#	DM	p-val
Best	1.35	0.02	1.86	N/A	1	0.78	0.22	0.83	0.11	0.69	N/A	1	2.12	0.02
90%	1.42	-0.07	2.03	N/A	1	-0.43	0.67	0.85	0.04	0.74	N/A	1	2.97	0.00
Median	1.43	0.12	2.06	N/A	1	-0.31	0.62	0.91	0.01	0.83	N/A	1	-4.08	1.00
10%	1.50	-0.07	2.28	N/A	1	-1.38	0.92	0.98	0.12	0.96	N/A	1	-4.64	1.00
Worst	1.64	-0.08	2.71	N/A	1	-1.53	0.94	1.01	-0.08	1.03	N/A	1	-2.32	0.99
Average	1.41	0.05	2.00	N/A	23	N/A	N/A	0.87	0.06	0.77	N/A	23	N/A	N/A

	Unemployment						
Regularization Group	RMSE	Bias	Var.	λ^*	#	DM	p-val
Ridge	0.69	0.00	0.48	4.87	23.00	-0.44	0.67
LASSO	0.69	0.00	0.48	0.05	5.27	-0.32	0.62
eRidge	0.66	0.05	0.44	max	23.00	5.33	0.00
eLASSO	0.67	0.09	0.45	5.66	23.00	51.57	0.00
peLASSO (Average)	0.63	-0.07	0.40	1.97	2.51	2.02	0.02
peLASSO (eRidge)	0.63	-0.04	0.40	(1.97, max)	2.51	2.25	0.01
peLASSO (eLASSO)	0.63	-0.04	0.39	(1.97, 6.58)	2.51	2.57	0.01
Comparisons	RMSE	Bias	Var.	λ^*	#	DM	p-val
Best	0.64	-0.14	0.40	N/A	1	1.16	0.12
90%	0.66	0.08	0.44	N/A	1	1.01	0.16
Median	0.71	0.11	0.51	N/A	1	-7.17	1.00
10%	0.80	0.07	0.65	N/A	1	-7.31	1.00
Worst	0.84	0.00	0.71	N/A	1	-2.76	1.00
Average	0.68	0.08	0.46	N/A	23	N/A	N/A

Table 13: **Combining methods forecast RMSEs based on ex post optimal λs.** λ^* is the ex post optimal penalty parameter, # is the average number of forecasters selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

	Real GDP							*Inflation*						
Regularization Group	RMSE	Bias	Var.	λ^*	#	DM	p-val	RMSE	Bias	Var.	λ^*	#	DM	p-val
Ridge	1.51	-0.03	2.31	115.44	23.00	-4.98	1.00	0.92	-0.03	0.85	85.39	23.00	-2.71	1.00
LASSO	1.72	0.15	2.97	0.59	2.29	-1.48	0.93	0.96	0.14	0.91	0.38	2.29	-3.38	1.00
eRidge	1.41	0.03	2.02	max	23.00	-13.18	1.00	0.86	0.05	0.75	max	23.00	12.41	0.00
eLASSO	1.41	0.03	2.01	3.60	23.00	3.19	0.00	0.86	0.04	0.75	1.08	23.00	14.85	0.00
peLASSO (Average)	1.40	0.11	1.98	0.11	4.60	0.39	0.35	0.84	0.08	0.71	0.02	7.71	10.65	0.00
peLASSO (eRidge)	1.40	0.11	1.98	(0.11, max)	4.60	0.39	0.35	0.84	0.08	0.71	(0.02, max)	7.71	10.65	0.00
peLASSO (eLASSO)	1.40	0.11	1.98	(0.11, 4.19)	4.60	0.39	0.35	0.84	0.08	0.71	(0.02, 1.25)	7.71	10.65	0.00
Comparisons	RMSE	Bias	Var.	λ^*	#	DM	p-val	RMSE	Bias	Var.	λ^*	#	DM	p-val
Best	1.36	0.01	1.87	N/A	1	0.78	0.22	0.83	0.11	0.69	N/A	1	1.36	0.09
90%	1.41	0.02	2.00	N/A	1	0.54	0.29	0.85	0.12	0.73	N/A	1	2.11	0.02
Median	1.45	0.11	2.12	N/A	1	-0.63	0.74	0.90	0.07	0.81	N/A	1	-6.25	1.00
10%	1.54	-0.01	2.40	N/A	1	-0.83	0.80	0.98	0.28	0.89	N/A	1	-5.63	1.00
Worst	1.63	-0.09	2.70	N/A	1	-1.45	0.93	1.01	0.19	1.00	N/A	1	-4.13	1.00
Average	1.41	0.03	2.01	N/A	23	N/A	N/A	0.86	0.06	0.75	N/A	23	N/A	N/A

	Unemployment						
Regularization Group	RMSE	Bias	Var.	λ^*	#	DM	p-val
Ridge	0.85	0.09	0.73	73.44	23.00	-8.72	1.00
LASSO	0.88	0.10	0.77	0.44	3.39	-4.99	1.00
eRidge	0.69	0.08	0.47	max	23.00	-40.33	1.00
eLASSO	0.68	0.10	0.46	8.90	23.00	-1.62	0.95
peLASSO (Average)	0.67	0.07	0.45	0.01	14.94	10.41	0.00
peLASSO (eRidge)	0.74	0.01	0.56	(0.01, max)	14.94	-5.60	1.00
peLASSO (eLASSO)	0.74	0.03	0.55	(0.01, 8.90)	14.94	-4.95	1.00
Comparisons	RMSE	Bias	Var.	λ^*	#	DM	p-val
Best	0.64	-0.14	0.40	N/A	1	1.52	0.06
90%	0.68	0.08	0.47	N/A	1	-0.16	0.56
Median	0.72	0.09	0.51	N/A	1	-6.06	1.00
10%	0.82	0.33	0.57	N/A	1	-4.75	1.00
Worst	0.94	0.24	0.83	N/A	1	-2.85	1.00
Average	0.68	0.08	0.46	N/A	23	N/A	N/A

Table 14: **Combining methods forecast RMSEs based on ex post optimal λs (real-time applicable).** λ^* is the ex post optimal penalty parameter, # is the average number of forecasters selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

A.2 Average-best forecast combination

	Real GDP						*Inflation*					
Average-best N	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	1.38	0.13	1.90	1.00	0.25	0.40	0.84	0.15	0.68	1.00	2.32	0.01
$N = 2$	1.35	0.11	1.81	2.00	1.62	0.05	0.81	0.12	0.64	2.00	3.95	0.00
$N = 3$	1.37	0.11	1.88	3.00	0.64	0.26	0.82	0.09	0.66	3.00	6.25	0.00
$N = 4$	1.36	0.11	1.83	4.00	1.42	0.08	0.82	0.09	0.67	4.00	7.35	0.00
$N = 5$	1.37	0.12	1.86	5.00	1.17	0.12	0.84	0.09	0.70	5.00	7.76	0.00
$N = 6$	1.37	0.11	1.87	6.00	1.30	0.10	0.85	0.08	0.71	6.00	10.20	0.00
Average-best $\leq N_{max}$	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
$N_{max} = 1$	1.38	0.13	1.90	1.00	0.25	0.40	0.84	0.15	0.68	1.00	2.32	0.01
$N_{max} = 2$	1.39	0.10	1.91	1.99	0.45	0.33	0.81	0.12	0.64	2.00	3.95	0.00
$N_{max} = 3$	1.40	0.10	1.96	2.44	0.01	0.49	0.83	0.12	0.67	2.10	4.18	0.00
$N_{max} = 4$	1.41	0.10	1.97	3.13	0.00	0.50	0.83	0.12	0.67	2.10	4.18	0.00
$N_{max} = 5$	1.41	0.09	1.99	3.46	-0.17	0.57	0.83	0.12	0.67	2.10	4.18	0.00
$N_{max} = 6$	1.41	0.09	1.99	3.46	-0.17	0.57	0.83	0.12	0.67	2.10	4.18	0.00
Comparisons	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
Best	1.35	0.02	1.86	1	0.78	0.22	0.83	0.11	0.69	1	2.12	0.02
90%	1.42	-0.07	2.03	1	-0.43	0.67	0.85	0.04	0.74	1	2.97	0.00
Median	1.43	0.12	2.06	1	-0.31	0.62	0.91	0.01	0.83	1	-4.08	1.00
10%	1.50	-0.07	2.28	1	-1.38	0.92	0.98	0.12	0.96	1	-4.64	1.00
Worst	1.64	-0.08	2.71	1	-1.53	0.94	1.01	-0.08	1.03	1	-2.32	0.99
Average	1.41	0.05	2.00	23	N/A	N/A	0.87	0.06	0.77	23	N/A	N/A

	Unemployment					
Average-best N	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	0.72	-0.04	0.52	1.00	-1.59	0.94
$N = 2$	0.66	-0.04	0.43	2.00	1.24	0.11
$N = 3$	0.67	-0.03	0.44	3.00	0.89	0.19
$N = 4$	0.67	-0.01	0.46	4.00	0.10	0.46
$N = 5$	0.67	0.00	0.44	5.00	1.53	0.06
$N = 6$	0.66	0.02	0.44	6.00	3.31	0.00
Average-best $\leq N_{max}$	RMSE	Bias	Var.	#	DM	p-val
$N_{max} = 1$	0.72	-0.04	0.52	1.00	-1.59	0.94
$N_{max} = 2$	0.66	-0.05	0.44	1.86	0.70	0.24
$N_{max} = 3$	0.66	-0.05	0.44	1.89	0.64	0.26
$N_{max} = 4$	0.66	-0.05	0.44	1.89	0.64	0.26
$N_{max} = 5$	0.66	-0.05	0.44	1.89	0.64	0.26
$N_{max} = 6$	0.66	-0.05	0.44	1.89	0.64	0.26
Comparisons	RMSE	Bias	Var.	#	DM	p-val
Best	0.64	-0.14	0.40	1	1.16	0.12
90%	0.66	0.08	0.44	1	1.01	0.16
Median	0.71	0.11	0.51	1	-7.17	1.00
10%	0.80	0.07	0.65	1	-7.31	1.00
Worst	0.84	0.00	0.71	1	-2.76	1.00
Average	0.68	0.08	0.46	23	N/A	N/A

Table 15: **Individual-based average-best forecast combination.** # is the average number of forecasters selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

	Real GDP							*Inflation*						
Average-best ($\leq 6, W$)	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	1.30	0.07	1.70	2.64	1	1.11	0.13	0.82	0.06	0.67	2.50	1	9.71	0.00
$W = 2$	1.29	0.07	1.65	2.66	2	1.86	0.03	0.79	0.09	0.62	2.93	2	13.05	0.00
$W = 3$	1.30	0.11	1.67	1.93	3	1.62	0.05	0.81	0.06	0.66	3.06	3	17.20	0.00
$W = 4$	1.35	0.09	1.82	3.36	4	1.14	0.13	0.81	0.07	0.66	2.11	4	16.01	0.00
$W = 5$	1.30	0.09	1.68	3.26	5	3.07	0.00	0.80	0.06	0.64	1.77	5	8.74	0.00
$W = 6$	1.33	0.11	1.75	2.57	6	1.67	0.05	0.79	0.08	0.61	2.00	6	5.75	0.00
$W = 7$	1.34	0.09	1.79	2.66	7	0.92	0.18	0.79	0.09	0.61	2.00	7	5.14	0.00
$W = 8$	1.37	0.08	1.86	3.00	8	0.73	0.23	0.79	0.08	0.62	2.00	8	5.18	0.00
$W = 9$	1.37	0.09	1.88	2.79	9	0.47	0.32	0.80	0.08	0.64	1.96	9	8.69	0.00
$W = 10$	1.38	0.08	1.89	3.63	10	0.51	0.30	0.79	0.07	0.62	1.90	10	11.18	0.00
$W = 15$	1.40	0.10	1.94	3.37	15	0.13	0.45	0.83	0.12	0.68	2.11	15	2.34	0.01
$W = 20$	1.41	0.09	1.97	3.66	20	-0.03	0.51	0.83	0.12	0.67	2.27	20	4.11	0.00
$W = 25$	1.42	0.11	2.01	2.61	25	-0.35	0.64	0.82	0.11	0.66	2.49	25	4.85	0.00
$W = 30$	1.42	0.09	2.00	2.46	30	-0.14	0.56	0.83	0.11	0.68	2.00	30	3.36	0.00
$W = 35$	1.45	0.14	2.08	2.53	35	-0.35	0.64	0.85	0.07	0.72	3.46	35	5.83	0.00
$W = 40$	1.45	0.13	2.09	2.53	40	-0.39	0.65	0.84	0.06	0.71	2.86	40	12.27	0.00
Average-best ($\leq 6, \leq 40$)	1.29	0.08	1.68	2.40	2.27	1.78	0.04	0.81	0.07	0.66	2.07	8.76	8.07	0.00
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	1.35	0.02	1.86	1	N/A	0.78	0.22	0.83	0.11	0.69	1	N/A	2.12	0.02
90%	1.42	-0.07	2.03	1	N/A	-0.43	0.67	0.85	0.04	0.74	1	N/A	2.97	0.00
Median	1.43	0.12	2.06	1	N/A	-0.31	0.62	0.91	0.01	0.83	1	N/A	-4.08	1.00
10%	1.50	-0.07	2.28	1	N/A	-1.38	0.92	0.98	0.12	0.96	1	N/A	-4.64	1.00
Worst	1.64	-0.08	2.71	1	N/A	-1.53	0.94	1.01	-0.08	1.03	1	N/A	-2.32	0.99
Average	1.41	0.05	2.00	23	N/A	N/A	N/A	0.87	0.06	0.77	23	N/A	N/A	N/A

	Unemployment						
Average-best ($\leq 6, W$)	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	0.56	0.09	0.30	2.61	1	17.81	0.00
$W = 2$	0.62	0.11	0.37	2.44	2	3.67	0.00
$W = 3$	0.60	0.06	0.36	2.09	3	4.54	0.00
$W = 4$	0.60	0.05	0.35	2.21	4	5.19	0.00
$W = 5$	0.63	0.07	0.39	1.39	5	2.38	0.01
$W = 6$	0.63	0.03	0.40	1.37	6	2.46	0.01
$W = 7$	0.63	0.01	0.40	1.61	7	2.39	0.01
$W = 8$	0.65	-0.01	0.42	2.16	8	1.34	0.09
$W = 9$	0.65	-0.04	0.42	1.93	9	1.03	0.15
$W = 10$	0.66	-0.01	0.43	2.29	10	0.87	0.19
$W = 15$	0.67	-0.01	0.45	3.36	15	0.34	0.37
$W = 20$	0.68	-0.03	0.46	2.03	20	-0.25	0.60
$W = 25$	0.67	-0.02	0.45	1.83	25	0.04	0.49
$W = 30$	0.68	-0.01	0.47	2.17	30	-0.53	0.70
$W = 35$	0.69	-0.01	0.48	2.84	35	-1.52	0.94
$W = 40$	0.67	0.02	0.45	3.29	40	0.26	0.40
Average-best ($\leq 6, \leq 40$)	0.58	0.03	0.35	1.61	4.73	4.79	0.00
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	0.64	-0.14	0.40	1	N/A	1.16	0.12
90%	0.66	0.08	0.44	1	N/A	1.01	0.16
Median	0.71	0.11	0.51	1	N/A	-7.17	1.00
10%	0.80	0.07	0.65	1	N/A	-7.31	1.00
Worst	0.84	0.00	0.71	1	N/A	-2.76	1.00
Average	0.68	0.08	0.46	23	N/A	N/A	N/A

Table 16: **Forecast combination.** $\#N$ is the average number of forecasters selected, $\#W$ is the average window width selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

	Real GDP						*Inflation*					
Average-best N	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	1.57	0.16	2.39	1.00	-0.78	0.78	0.87	0.10	0.76	1.00	-0.63	0.73
$N = 3$	1.47	0.13	2.09	3.00	-1.11	0.87	0.85	0.07	0.72	3.00	1.81	0.04
$N = 5$	1.46	0.10	2.05	5.00	-2.65	1.00	0.84	0.05	0.70	5.00	6.09	0.00
$N = 7$	1.45	0.09	2.04	7.00	-4.97	1.00	0.84	0.05	0.70	7.00	10.52	0.00
$N = 10$	1.45	0.08	2.03	10.00	-5.93	1.00	0.86	0.05	0.74	10.00	3.64	0.00
$N = 15$	1.43	0.05	1.99	15.00	-2.76	1.00	0.86	0.05	0.74	15.00	5.94	0.00
$N = 20$	1.43	0.03	1.99	20.00	-1.79	0.96	0.86	0.05	0.73	20.00	36.10	0.00
Average-best *all*	1.48	0.09	2.19	10.67	-1.33	0.91	0.85	0.06	0.72	2.77	2.39	0.01
Comparisons	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
Best	1.36	0.01	1.87	1	0.78	0.22	0.83	0.11	0.69	1	1.36	0.09
90%	1.41	0.02	2.00	1	0.54	0.29	0.85	0.12	0.73	1	2.11	0.02
Median	1.45	0.11	2.12	1	-0.63	0.74	0.90	0.07	0.81	1	-6.25	1.00
10%	1.54	-0.01	2.40	1	-0.83	0.80	0.98	0.28	0.89	1	-5.63	1.00
Worst	1.63	-0.09	2.70	1	-1.45	0.93	1.01	0.19	1.00	1	-4.13	1.00
Average	1.41	0.03	2.01	23	N/A	N/A	0.86	0.06	0.75	23	N/A	N/A

	Unemployment					
Average-best N	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	0.76	-0.03	0.56	1.00	-2.69	1.00
$N = 3$	0.73	0.03	0.53	3.00	-7.70	1.00
$N = 5$	0.71	0.05	0.51	5.00	-13.32	1.00
$N = 7$	0.72	0.06	0.52	7.00	-17.00	1.00
$N = 10$	0.70	0.08	0.49	10.00	-28.40	1.00
$N = 15$	0.69	0.08	0.47	15.00	-17.79	1.00
$N = 20$	0.68	0.08	0.46	20.00	-12.16	1.00
Average-best *all*	0.72	0.05	0.52	10.89	-5.31	1.00
Comparisons	RMSE	Bias	Var.	#	DM	p-val
Best	0.64	-0.14	0.40	1	1.52	0.06
90%	0.68	0.08	0.47	1	-0.16	0.56
Median	0.72	0.09	0.51	1	-6.06	1.00
10%	0.82	0.33	0.57	1	-4.75	1.00
Worst	0.94	0.24	0.83	1	-2.85	1.00
Average	0.68	0.08	0.46	23	N/A	N/A

Table 17: **Individual-based average-best forecast combination (real-time applicable).** # is the average number of forecasters selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

	Real GDP							*Inflation*						
Average-best (*all*, W)	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	1.45	0.05	2.10	14.13	1	-3.18	1.00	0.85	0.04	0.72	6.51	1	7.38	0.00
$W = 2$	1.42	0.05	2.02	14.21	2	-1.12	0.87	0.85	0.06	0.72	5.73	2	1.44	0.07
$W = 3$	1.46	0.05	2.13	8.59	3	-2.32	0.99	0.86	0.04	0.74	5.14	3	1.53	0.06
$W = 4$	1.48	0.13	2.16	5.94	4	-0.54	0.71	0.85	0.07	0.72	3.47	4	0.83	0.20
$W = 5$	1.43	0.07	2.05	4.43	5	-1.87	0.97	0.83	0.08	0.69	1.53	5	3.20	0.00
$W = 6$	1.43	0.10	2.03	5.57	6	-0.67	0.75	0.81	0.08	0.65	1.54	6	3.14	0.00
$W = 7$	1.44	0.10	2.07	13.96	7	-0.77	0.78	0.83	0.11	0.68	1.63	7	1.65	0.05
$W = 8$	1.43	0.07	2.05	11.03	8	-0.98	0.84	0.84	0.11	0.69	1.64	8	1.32	0.09
$W = 9$	1.42	0.10	2.02	9.24	9	-0.35	0.64	0.86	0.09	0.73	2.24	9	0.24	0.40
$W = 10$	1.42	0.07	2.02	9.41	10	-0.35	0.64	0.85	0.08	0.72	2.74	10	0.73	0.23
$W = 15$	1.44	0.09	2.06	8.77	15	-0.58	0.72	0.85	0.08	0.71	2.89	15	1.81	0.03
$W = 20$	1.50	0.11	2.24	10.37	20	-0.68	0.75	0.86	0.06	0.73	2.57	20	0.50	0.31
$W = 25$	1.43	0.11	2.04	7.51	25	-0.78	0.78	0.87	0.02	0.75	3.39	25	-0.25	0.60
$W = 30$	1.47	0.14	2.15	12.03	30	-0.39	0.65	0.85	-0.01	0.73	4.40	30	3.36	0.00
$W = 35$	1.47	0.14	2.16	12.03	35	-0.39	0.65	0.85	-0.01	0.73	4.66	35	3.31	0.00
$W = 40$	1.47	0.14	2.16	12.03	40	-0.40	0.65	0.86	-0.02	0.73	4.60	40	2.13	0.02
Average-best (*all*, ≤ 40)	1.47	0.08	2.18	2.43	6.64	-0.60	0.73	0.83	0.07	0.69	1.54	6.50	2.19	0.01
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	1.36	0.01	1.87	1	N/A	0.78	0.22	0.83	0.11	0.69	1	N/A	1.36	0.09
90%	1.41	0.02	2.00	1	N/A	0.54	0.29	0.85	0.12	0.73	1	N/A	2.11	0.02
Median	1.45	0.11	2.12	1	N/A	-0.63	0.74	0.90	0.07	0.81	1	N/A	-6.25	1.00
10%	1.54	-0.01	2.40	1	N/A	-0.83	0.80	0.98	0.28	0.89	1	N/A	-5.63	1.00
Worst	1.63	-0.09	2.70	1	N/A	-1.45	0.93	1.01	0.19	1.00	1	N/A	-4.13	1.00
Average	1.41	0.03	2.01	23	N/A	N/A	N/A	0.86	0.06	0.75	23	N/A	N/A	N/A

	Unemployment						
Average-best (*all*, W)	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	0.70	0.04	0.49	8.23	1	-2.60	1.00
$W = 2$	0.71	0.08	0.49	8.76	2	-21.35	1.00
$W = 3$	0.70	0.05	0.49	10.81	3	-2.98	1.00
$W = 4$	0.71	0.04	0.51	10.59	4	-4.92	1.00
$W = 5$	0.71	0.03	0.51	8.77	5	-2.62	1.00
$W = 6$	0.70	0.00	0.50	9.53	6	-1.71	0.96
$W = 7$	0.71	-0.01	0.50	8.90	7	-1.43	0.92
$W = 8$	0.73	0.03	0.53	10.57	8	-3.41	1.00
$W = 9$	0.72	0.02	0.52	10.84	9	-2.72	1.00
$W = 10$	0.72	0.01	0.52	10.71	10	-3.08	1.00
$W = 15$	0.69	0.03	0.47	10.46	15	-1.80	0.96
$W = 20$	0.71	0.05	0.51	12.23	20	-5.27	1.00
$W = 25$	0.69	0.04	0.47	8.76	25	-2.50	0.99
$W = 30$	0.70	0.08	0.48	12.90	30	-3.89	1.00
$W = 35$	0.71	0.07	0.49	13.00	35	-4.76	1.00
$W = 40$	0.71	0.07	0.49	12.73	40	-4.91	1.00
Average-best (*all*, ≤ 40)	0.69	0.00	0.48	4.79	5.90	-0.53	0.70
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	0.64	-0.14	0.40	1	N/A	1.52	0.06
90%	0.68	0.08	0.47	1	N/A	-0.16	0.56
Median	0.72	0.09	0.51	1	N/A	-6.06	1.00
10%	0.82	0.33	0.57	1	N/A	-4.75	1.00
Worst	0.94	0.24	0.83	1	N/A	-2.85	1.00
Average	0.68	0.08	0.46	23	N/A	N/A	N/A

Table 18: **Forecast combination (real-time applicable).** #N is the average number of forecasters selected, #W is the average window width selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

A.3 Average-best density forecast combination

	Real GDP						*Inflation*					
Average-best N	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	1.46	0.14	2.11	1.00	-0.21	0.58	0.82	0.04	0.66	1.00	1.34	0.09
$N = 2$	1.38	0.11	1.89	2.00	0.81	0.21	0.80	0.05	0.63	2.00	3.50	0.00
$N = 3$	1.39	0.10	1.91	3.00	0.93	0.18	0.81	0.02	0.66	3.00	4.87	0.00
$N = 4$	1.38	0.07	1.90	4.00	1.60	0.05	0.82	0.03	0.68	4.00	5.86	0.00
$N = 5$	1.39	0.05	1.93	5.00	1.99	0.02	0.83	0.04	0.68	5.00	8.09	0.00
$N = 6$	1.38	0.05	1.90	6.00	2.50	0.01	0.82	0.05	0.67	6.00	9.09	0.00
Average-best $\leq N_{max}$	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
$N_{max} = 1$	1.46	0.14	2.11	1.00	-0.21	0.58	0.82	0.04	0.66	1.00	1.34	0.09
$N_{max} = 2$	1.40	0.10	1.95	1.90	0.17	0.43	0.80	0.07	0.64	1.84	1.98	0.02
$N_{max} = 3$	1.40	0.10	1.95	1.90	0.17	0.43	0.80	0.07	0.64	2.11	1.90	0.03
$N_{max} = 4$	1.41	0.10	1.96	1.96	0.10	0.46	0.80	0.07	0.64	2.11	1.90	0.03
$N_{max} = 5$	1.41	0.09	1.97	2.06	0.09	0.47	0.80	0.07	0.64	2.11	1.90	0.03
$N_{max} = 6$	1.41	0.09	1.97	2.19	0.10	0.46	0.80	0.07	0.64	2.11	1.90	0.03
Comparisons	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
Best	1.36	0.01	1.87	1	0.73	0.23	0.83	0.11	0.69	1	1.74	0.04
90%	1.40	-0.02	1.98	1	2.08	0.02	0.85	0.12	0.73	1	3.55	0.00
Median	1.44	-0.02	2.11	1	-5.74	1.00	0.90	0.12	0.80	1	-13.80	1.00
10%	1.52	-0.09	2.33	1	-2.56	0.99	0.96	0.07	0.94	1	-0.89	0.81
Worst	1.63	-0.09	2.70	1	-1.58	0.94	1.03	0.20	1.03	1	-4.06	1.00
Average	1.41	0.00	2.03	23	N/A	N/A	0.87	0.05	0.76	23	N/A	N/A

	Unemployment					
Average-best N	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	0.67	-0.02	0.44	1.00	1.47	0.07
$N = 2$	0.67	-0.01	0.45	2.00	1.72	0.04
$N = 3$	0.66	0.03	0.44	3.00	5.23	0.00
$N = 4$	0.68	0.05	0.45	4.00	5.68	0.00
$N = 5$	0.68	0.07	0.45	5.00	8.29	0.00
$N = 6$	0.68	0.06	0.45	6.00	11.18	0.00
Average-best $\leq N_{max}$	RMSE	Bias	Var.	#	DM	p-val
$N_{max} = 1$	0.67	-0.02	0.44	1.00	1.47	0.07
$N_{max} = 2$	0.68	-0.01	0.46	1.60	1.11	0.13
$N_{max} = 3$	0.68	-0.01	0.46	1.87	1.26	0.10
$N_{max} = 4$	0.68	0.00	0.46	2.10	1.03	0.15
$N_{max} = 5$	0.68	0.00	0.46	2.10	1.03	0.15
$N_{max} = 6$	0.68	0.00	0.46	2.10	1.03	0.15
Comparisons	RMSE	Bias	Var.	#	DM	p-val
Best	0.64	-0.14	0.40	1	1.79	0.04
90%	0.68	0.09	0.46	1	2.88	0.00
Median	0.72	0.13	0.50	1	-3.06	1.00
10%	0.82	0.33	0.57	1	-4.56	1.00
Worst	0.95	0.26	0.85	1	-2.90	1.00
Average	0.70	0.09	0.48	23	N/A	N/A

Table 19: **Individual-based average-best density forecast combination.** # is the average number of forecasters selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

	Real GDP							*Inflation*						
Average-best ($\leq 6, W$)	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	1.36	0.04	1.85	3.31	1	0.57	0.28	0.82	0.00	0.66	3.00	1	15.82	0.00
$W = 2$	1.35	0.07	1.82	2.04	2	0.46	0.32	0.81	0.01	0.65	2.74	2	12.73	0.00
$W = 3$	1.38	0.02	1.90	3.21	3	0.90	0.18	0.81	0.04	0.66	3.57	3	26.12	0.00
$W = 4$	1.38	0.07	1.90	2.44	4	0.23	0.41	0.83	0.06	0.69	1.67	4	2.35	0.01
$W = 5$	1.31	0.03	1.72	2.83	5	1.60	0.05	0.83	0.06	0.68	1.64	5	1.52	0.06
$W = 6$	1.37	0.01	1.87	2.99	6	0.57	0.28	0.81	0.10	0.65	1.56	6	2.02	0.02
$W = 7$	1.36	0.04	1.86	2.67	7	0.63	0.27	0.83	0.08	0.68	1.90	7	1.51	0.07
$W = 8$	1.36	0.06	1.85	2.20	8	0.77	0.22	0.83	0.09	0.68	1.43	8	1.20	0.12
$W = 9$	1.33	0.05	1.76	2.07	9	1.32	0.09	0.83	0.10	0.69	1.30	9	1.00	0.16
$W = 10$	1.40	0.04	1.96	2.07	10	0.25	0.40	0.82	0.08	0.67	1.63	10	2.14	0.02
$W = 15$	1.40	0.07	1.97	4.06	15	0.21	0.42	0.83	0.05	0.69	1.80	15	1.65	0.05
$W = 20$	1.41	0.10	1.98	2.13	20	0.06	0.48	0.80	0.06	0.64	1.76	20	1.98	0.02
$W = 25$	1.40	0.15	1.94	2.06	25	0.15	0.44	0.81	0.06	0.66	1.77	25	1.70	0.04
$W = 30$	1.41	0.16	1.98	2.09	30	-0.01	0.51	0.84	0.03	0.70	2.87	30	2.88	0.00
$W = 35$	1.40	0.15	1.94	2.04	35	0.15	0.44	0.86	0.00	0.73	3.33	35	3.10	0.00
$W = 40$	1.43	0.13	2.03	2.40	40	-0.19	0.58	0.85	-0.01	0.73	3.19	40	3.90	0.00
Average-best ($\leq 6, \leq 40$)	1.38	0.06	1.92	1.64	2.34	0.27	0.39	0.83	0.09	0.69	2.19	8.13	1.34	0.09
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	1.36	0.01	1.87	1	N/A	0.73	0.23	0.83	0.11	0.69	1	N/A	1.74	0.04
90%	1.40	-0.02	1.98	1	N/A	2.08	0.02	0.85	0.12	0.73	1	N/A	3.55	0.00
Median	1.44	-0.02	2.11	1	N/A	-5.74	1.00	0.90	0.12	0.80	1	N/A	-13.80	1.00
10%	1.52	-0.09	2.33	1	N/A	-2.56	0.99	0.96	0.07	0.94	1	N/A	-0.89	0.81
Worst	1.63	-0.09	2.70	1	N/A	-1.58	0.94	1.03	0.20	1.03	1	N/A	-4.06	1.00
Average	1.41	0.00	2.03	23	N/A	N/A	N/A	0.87	0.05	0.76	23	N/A	N/A	N/A

Unemployment

Average-best ($\leq 6, W$)	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	0.61	0.03	0.37	2.30	1	7.95	0.00
$W = 2$	0.62	0.06	0.38	1.74	2	6.47	0.00
$W = 3$	0.61	0.07	0.37	1.99	3	12.25	0.00
$W = 4$	0.61	0.09	0.36	3.20	4	16.04	0.00
$W = 5$	0.63	0.08	0.39	2.56	5	10.98	0.00
$W = 6$	0.65	0.09	0.42	2.33	6	7.10	0.00
$W = 7$	0.66	0.05	0.44	2.33	7	4.77	0.00
$W = 8$	0.65	0.04	0.42	2.29	8	4.85	0.00
$W = 9$	0.66	0.02	0.43	2.31	9	4.00	0.00
$W = 10$	0.67	0.03	0.44	3.77	10	6.49	0.00
$W = 15$	0.67	0.03	0.45	3.11	15	2.33	0.01
$W = 20$	0.69	0.01	0.48	2.34	20	0.23	0.41
$W = 25$	0.65	0.03	0.42	2.69	25	4.46	0.00
$W = 30$	0.67	0.05	0.44	3.97	30	10.00	0.00
$W = 35$	0.70	0.07	0.48	4.11	35	-1.51	0.93
$W = 40$	0.71	0.08	0.49	4.81	40	-4.09	1.00
Average-best ($\leq 6, \leq 40$)	0.64	0.03	0.41	1.90	3.64	4.68	0.00
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	0.64	-0.14	0.40	1	N/A	1.79	0.04
90%	0.68	0.09	0.46	1	N/A	2.88	0.00
Median	0.72	0.13	0.50	1	N/A	-3.06	1.00
10%	0.82	0.33	0.57	1	N/A	-4.56	1.00
Worst	0.95	0.26	0.85	1	N/A	-2.90	1.00
Average	0.70	0.09	0.48	23	N/A	N/A	N/A

Table 20: **Density forecast combination.** $\#N$ is the average number of forecasters selected, $\#W$ is the average window width selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

	Real GDP						*Inflation*					
Average-best N	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	1.51	0.25	2.22	1.00	-0.45	0.68	0.89	0.03	0.78	1.00	-0.40	0.65
$N = 3$	1.42	0.11	2.01	3.00	-0.24	0.59	0.83	0.02	0.69	3.00	3.25	0.00
$N = 5$	1.42	0.07	2.02	5.00	-0.78	0.78	0.84	0.04	0.70	5.00	5.27	0.00
$N = 7$	1.41	0.05	1.97	7.00	0.92	0.18	0.84	0.05	0.71	7.00	9.01	0.00
$N = 10$	1.40	0.03	1.96	10.00	3.69	0.00	0.86	0.05	0.73	10.00	10.19	0.00
$N = 15$	1.41	0.03	1.99	15.00	1.04	0.15	0.86	0.05	0.74	15.00	20.62	0.00
$N = 20$	1.41	0.01	1.98	20.00	8.42	0.00	0.87	0.05	0.75	20.00	-0.21	0.58
Average-best *all*	1.47	0.20	2.11	5.29	-0.25	0.60	0.86	0.04	0.74	3.19	0.40	0.34
Comparisons	RMSE	Bias	Var.	#	DM	p-val	RMSE	Bias	Var.	#	DM	p-val
Best	1.36	0.01	1.87	1	0.73	0.23	0.83	0.11	0.69	1	1.74	0.04
90%	1.40	-0.02	1.98	1	2.08	0.02	0.85	0.12	0.73	1	3.55	0.00
Median	1.44	-0.02	2.11	1	-5.74	1.00	0.90	0.12	0.80	1	-13.80	1.00
10%	1.52	-0.09	2.33	1	-2.56	0.99	0.96	0.07	0.94	1	-0.89	0.81
Worst	1.63	-0.09	2.70	1	-1.58	0.94	1.03	0.20	1.03	1	-4.06	1.00
Average	1.41	0.00	2.03	23	N/A	N/A	0.87	0.05	0.76	23	N/A	N/A

	Unemployment					
Average-best N	RMSE	Bias	Var.	#	DM	p-val
$N = 1$	0.73	-0.03	0.53	1.00	-1.33	0.91
$N = 3$	0.68	0.05	0.47	3.00	2.82	0.00
$N = 5$	0.69	0.07	0.47	5.00	4.32	0.00
$N = 7$	0.68	0.07	0.46	7.00	10.46	0.00
$N = 10$	0.69	0.08	0.47	10.00	7.83	0.00
$N = 15$	0.68	0.08	0.45	15.00	49.79	0.00
$N = 20$	0.69	0.08	0.47	20.00	112.34	0.00
Average-best *all*	0.74	0.04	0.54	3.56	-3.43	1.00
Comparisons	RMSE	Bias	Var.	#	DM	p-val
Best	0.64	-0.14	0.40	1	1.79	0.04
90%	0.68	0.09	0.46	1	2.88	0.00
Median	0.72	0.13	0.50	1	-3.06	1.00
10%	0.82	0.33	0.57	1	-4.56	1.00
Worst	0.95	0.26	0.85	1	-2.90	1.00
Average	0.70	0.09	0.48	23	N/A	N/A

Table 21: **Individual-based average-best density forecast combination (real-time applicable).** # is the average number of forecasters selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).

	Real GDP							*Inflation*						
Average-best (*all*, W)	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	1.42	0.02	2.03	7.37	1	-0.76	0.78	0.87	0.04	0.76	2.33	1	-0.14	0.56
$W = 2$	1.45	0.02	2.10	6.19	2	-0.80	0.79	0.87	0.05	0.76	3.54	2	-0.51	0.69
$W = 3$	1.47	0.12	2.16	5.90	3	-0.34	0.63	0.85	0.05	0.73	3.93	3	2.19	0.01
$W = 4$	1.47	0.12	2.14	4.63	4	-0.59	0.72	0.86	0.08	0.74	4.87	4	1.14	0.13
$W = 5$	1.46	0.14	2.12	2.40	5	-0.26	0.60	0.87	0.10	0.75	3.54	5	-0.18	0.57
$W = 6$	1.46	0.17	2.11	3.61	6	-0.25	0.60	0.87	0.12	0.75	3.01	6	-0.16	0.56
$W = 7$	1.46	0.12	2.12	6.89	7	-0.36	0.64	0.86	0.11	0.72	1.86	7	0.27	0.40
$W = 8$	1.42	0.10	2.01	5.23	8	-0.18	0.57	0.85	0.11	0.71	1.83	8	0.57	0.29
$W = 9$	1.43	0.08	2.05	5.76	9	-0.55	0.71	0.85	0.10	0.71	1.64	9	0.49	0.31
$W = 10$	1.42	0.07	2.02	6.17	10	-0.23	0.59	0.87	0.07	0.74	2.24	10	0.15	0.44
$W = 15$	1.43	0.09	2.03	6.29	15	-0.42	0.66	0.83	0.05	0.69	2.30	15	2.03	0.02
$W = 20$	1.45	0.16	2.08	5.26	20	-0.19	0.58	0.86	0.04	0.73	2.41	20	0.60	0.28
$W = 25$	1.39	0.17	1.90	1.71	25	0.25	0.40	0.88	0.04	0.77	3.01	25	-0.70	0.76
$W = 30$	1.45	0.17	2.06	3.26	30	-0.17	0.57	0.87	0.02	0.76	6.21	30	-0.49	0.69
$W = 35$	1.45	0.16	2.07	5.63	35	-0.17	0.57	0.88	0.01	0.78	7.10	35	-2.73	1.00
$W = 40$	1.45	0.15	2.08	7.33	40	-0.18	0.57	0.87	-0.01	0.76	6.33	40	-1.26	0.90
Average-best (*all*, ≤ 40)	1.40	0.18	1.95	1.80	18.97	0.17	0.43	0.85	0.06	0.73	2.01	11.61	0.48	0.32
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	1.36	0.01	1.87	1	N/A	0.73	0.23	0.83	0.11	0.69	1	N/A	1.74	0.04
90%	1.40	-0.02	1.98	1	N/A	2.08	0.02	0.85	0.12	0.73	1	N/A	3.55	0.00
Median	1.44	-0.02	2.11	1	N/A	-5.74	1.00	0.90	0.12	0.80	1	N/A	-13.80	1.00
10%	1.52	-0.09	2.33	1	N/A	-2.56	0.99	0.96	0.07	0.94	1	N/A	-0.89	0.81
Worst	1.63	-0.09	2.70	1	N/A	-1.58	0.94	1.03	0.20	1.03	1	N/A	-4.06	1.00
Average	1.41	0.00	2.03	23	N/A	N/A	N/A	0.87	0.05	0.76	23	N/A	N/A	N/A

	Unemployment						
Average-best (*all*, W)	RMSE	Bias	Var.	#N	#W	DM	p-val
$W = 1$	0.72	0.06	0.51	13.80	1	-3.75	1.00
$W = 2$	0.69	0.09	0.47	16.43	2	22.67	0.00
$W = 3$	0.70	0.09	0.48	17.84	3	-5.25	1.00
$W = 4$	0.69	0.08	0.47	13.01	4	1.41	0.08
$W = 5$	0.69	0.07	0.48	11.69	5	1.84	0.03
$W = 6$	0.69	0.06	0.48	8.03	6	2.00	0.02
$W = 7$	0.70	0.07	0.49	7.87	7	-2.90	1.00
$W = 8$	0.71	0.04	0.51	10.79	8	-3.62	1.00
$W = 9$	0.70	0.07	0.48	11.10	9	-3.50	1.00
$W = 10$	0.72	0.05	0.51	11.34	10	-4.76	1.00
$W = 15$	0.68	0.01	0.47	5.56	15	0.65	0.26
$W = 20$	0.69	0.01	0.48	5.47	20	0.12	0.45
$W = 25$	0.68	0.02	0.47	3.40	25	1.15	0.13
$W = 30$	0.70	0.10	0.48	11.70	30	-1.45	0.93
$W = 35$	0.70	0.09	0.48	11.84	35	-0.92	0.82
$W = 40$	0.70	0.09	0.48	11.81	40	-0.76	0.78
Average-best (*all*, ≤ 40)	0.69	0.01	0.48	3.04	14.83	0.60	0.27
Comparisons	RMSE	Bias	Var.	#N	#W	DM	p-val
Best	0.64	-0.14	0.40	1	N/A	1.79	0.04
90%	0.68	0.09	0.46	1	N/A	2.88	0.00
Median	0.72	0.13	0.50	1	N/A	-3.06	1.00
10%	0.82	0.33	0.57	1	N/A	-4.56	1.00
Worst	0.95	0.26	0.85	1	N/A	-2.90	1.00
Average	0.70	0.09	0.48	23	N/A	N/A	N/A

Table 22: **Density forecast combination (real-time applicable).** $\#N$ is the average number of forecasters selected, $\#W$ is the average window width selected, DM is the one-sided (Diebold and Mariano 2002) statistic against a simple average, and p-val represents the p-value. We compute DM as per Harvey et al. (1997).