UNIVERSITY of York

Discussion Papers in Economics

No. 19/14

A Real-time Density Forecast Evaluation of the ECB Survey of Professional Forecasters

Laura Coroneo (University of York),
Fabrizio Iacone (Università degli Studi di Milano, University of York),
Fabio Profumo (University of York)

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

A Real-time Density Forecast Evaluation of the ECB Survey of Professional Forecasters

Laura Coroneo*
University of York

Fabrizio Iacone
Università degli Studi di Milano
University of York

Fabio Profumo
University of York

25th September 2019

Abstract

We evaluate the real-time predictive ability of density forecasts from the European Central Bank's Survey of Professional Forecasters (ECB SPF) using the Diebold and Mariano (1995) and West (1996) test. As the sample size for the ECB SPF is fairly small, we use fixed-b and fixed-m asymptotics to alleviate size distortions. We verify in an original Monte Carlo design that fixed-smoothing asymptotics delivers correctly sized tests in this framework. Empirical results indicate that ECB SPF density forecasts for unemployment and real GDP growth beat simple benchmarks at one-year horizon. ECB SPF density forecasts for inflation instead do not easily outperform simple benchmarks, as up to 2008 ECB SPF inflation expectations are close to the target. After 2008, we find that the predictive ability of the ECB SPF is more conspicuous for all variables, even though inflation expectations are still loosely anchored to the target.

Keywords: real-time density forecast evaluation, ECB SPF, Diebold-Mariano-West test, fixed-smoothing asymptotics

JEL Classification: C12, C22, E17

*Corresponding author: Laura Coroneo, Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, UK. The support of the ESRC grant ES/J500215/1 is gratefully acknowledged. We also thank Paulo Santos Monteiro for useful comments.

1 Introduction

Market expectations play a key role in determining the outcomes of monetary policies; for this reason, monetary authorities place a great emphasis on informing market expectations, for example using inflation targeting or forward guidance.

Measuring market expectations, therefore, is of crucial importance, and central banks around the world now regularly run surveys of professional forecasters to gather information about market expectations and the monetary policy stance. Survey respondents are asked to report their point forecasts for a set of macroeconomic fundamentals and, increasingly, to provide a density forecast that describes the predicted probability distribution of the variables of interest. Compared to the more popular point forecasts, density forecasts provide a wider understanding of the uncertainty associated with the prediction, see Fair (1980) and Dawid (1984) for some early references, and Tay and Wallis (2000) for a more recent detailed discussion.

Well-known examples of survey density forecasts include the Survey of Professional Forecasters (SPF) currently managed by the Federal Reserve Bank of Philadelphia, the Survey of External Forecasters managed by the Bank of England and the European Central Bank's Survey of Professional Forecasters (ECB SPF). A large amount of work has been devoted to analysing the density forecasts provided by the US SPF, see among others Diebold, Tay and Wallis (1997) and Clements (2014), and the Bank of England's Survey of External Forecasters, see among others Boero, Smith and Wallis (2008) and Mitchell and Hall (2005). As for the density forecasts provided by the ECB SPF, the literature is more limited because the survey started only in 1999.

In this paper, we conduct a real-time evaluation of the ECB SPF density forecasts for key macroeconomic variables. We are interested in establishing whether ECB SPF density forecasts can beat simple benchmarks for real GDP growth and unemployment, as the superior predictive ability of the forecasters would indicate that they use more sophisticated models to process the available information. On the contrary, if inflation

expectations are well-anchored to the target, we do not expect the professional forecasters to outperform simple benchmarks, see among others Geraats (2008), Hartmann and Smets (2018) and Henckel, Menzies, Moffatt and Zizzo (2019).

ECB SPF density forecasts are all reported as histograms, as forecasters report probabilities for intervals, or bins. Two loss functions that naturally accommodate this format of data are the Quadratic Probability Score by Brier (1950) and the Ranked Probability Score by Epstein (1969). We compare ECB SPF density forecasts to competing forecasts from simple models that should be easily beaten by professional forecasters: a uniform distribution assigning the same probability to all bins; a Gaussian distribution based on the assumption that the target variable follows a random walk without drift; and a naive forecast taken from the previous round of ECB SPF forecasts. All forecasts are produced in real-time, by using the same information available to professional forecasters at each survey deadline.

We test the null of equal predictive accuracy of the ECB SPF and the benchmarks using the Diebold and Mariano (1995) and West (1996) (DMW) test. The DMW framework is simple and the test statistic is easy to compute; these features also make it widely used for density forecast evaluation. However, as recognised by Diebold and Mariano (1995) themselves and documented by Clark (1999), it suffers from relevant small sample size distortion, that could lead to spurious results: this is obviously a concern in our case as we only have 62 observations. We, therefore, assess the predictive ability of the ECB SPF density forecasts using an alternative approach based on fixed-smoothing asymptotics. In particular, we consider fixed-b asymptotics by Kiefer and Vogelsang (2005) and fixed-m asymptotics by Hualde and Iacone (2017). This approach proved capable of eliminating size distortion in DMW test for comparing point forecasts, see Coroneo and Iacone (2015). In an original Monte Carlo exercise, we verify that fixed-smoothing asymptotics may also be used with success to evaluate density forecasts using the DMW test.

Our empirical results indicate that the ECB SPF density forecasts for unemployment and real GDP growth beat simple benchmarks at one-year horizon. The ECB SPF density forecasts for inflation instead do not easily outperform simple benchmarks, as up to 2008 ECB SPF inflation expectations and realisations are close to the target. After 2008, we find that the predictive ability of the ECB SPF is more conspicuous for all variables, even though inflation expectations are still loosely anchored to the target. Overall, we interpret these results as evidence that professional forecasters use more sophisticated models to predict the variables of interest, but they do not entirely ignore the information contained in the inflation target.

This paper is related to the growing literature on evaluating the ECB SPF, for a survey see de Vincent-Humphreys, Dimitrova, Falck and Henkel (2019). Our contribution is to assess the predictive ability of ECB SPF density forecasts with respect to simple benchmarks, as in Kenny, Kostka and Masera (2014), but by applying fixed-smoothing asymptotics we avoid the risk of spurious results. In addition, our results support the anecdotal evidence of a change in the forecasting practice after the financial crisis. In line with the empirical investigations of Lyziak and Paloviita (2017) and Grishchenko, Mouabbi and Renne (2019), we find that, after 2008, inflation is more loosely anchored to the target, and professional forecasters take this fact into account. Even in this phase, however, the outperformance of the ECB SPF density forecast for inflation is weaker than for unemployment and real GDP growth, signalling that inflation targeting still informs the forecasters.

The remainder of the paper proceeds as follows. In Section 2 we describe the structure of the ECB SPF survey, and in Section 3 we discuss revision in macro variables. In Section 4 we present our benchmark forecasts, and in Section 5 we describe the procedure for density forecast evaluation in detail, including a discussion of the DMW test with fixed-smoothing asymptotics. We investigate the properties of the test in Section 6, where we present a Monte Carlo exercise and provide recommendations for bandwidths.

In Section 7 we present the outcome of the ECB SPF density forecast evaluation, and in Section 8 we conclude.

2 The ECB Survey of Professional Forecasters

The ECB started the survey in 1999 to gather information about private sector expectations and to assess the credibility of the newly established central bank.

The ECB SPF is submitted quarterly to a panel of forecasters (about 80 institutions with an average of 60 responses each round). Participants are experts affiliated with financial or non-financial institutions based within the European Union, and form an heterogeneous group to guarantee the representativeness and independence of the expectations collected.

The ECB asks panelists to provide point and density forecasts for four main economic indicators for the euro area:

1. Inflation: defined as the year-on-year percentage change of the Harmonised Index of Consumer Prices (HICP) published by Eurostat;

2. Core inflation: defined as the year-on-year percentage change in the euro area HICP special aggregate "all items excluding energy, food, alcohol and tobacco" published by Eurostat (only available from the 2016.Q4 survey round);

3. Real GDP growth: defined as the year-on-year percentage change of real GDP, based on the standardised European System of National and Regional Accounts 2010 definition;

4. Unemployment rate: the unemployment rate refers to the International Labour Organization's definition and it is calculated as percentage of the labour force.

Between 1999.Q1 and 2001.Q3, the survey was conducted in the middle month of each quarter, i.e. in February, May, August and November; since 2001.Q4, the survey

Table 1: ECB SPF timing

Inflation				
Survey date	Deadline	Latest info available	Forecast 1 year	Forecast 2 years
Q1.Y	m1.Y	m12.Y-1	m12.Y	m12.Y+1
Q2.Y	m4.Y	m3.Y	m3.Y+1	m3.Y+2
Q3.Y	m7.Y	m6.Y	m6.Y+1	m6.Y+2
Q4.Y	m10.Y	m9.Y	m9.Y+1	m9.Y+2

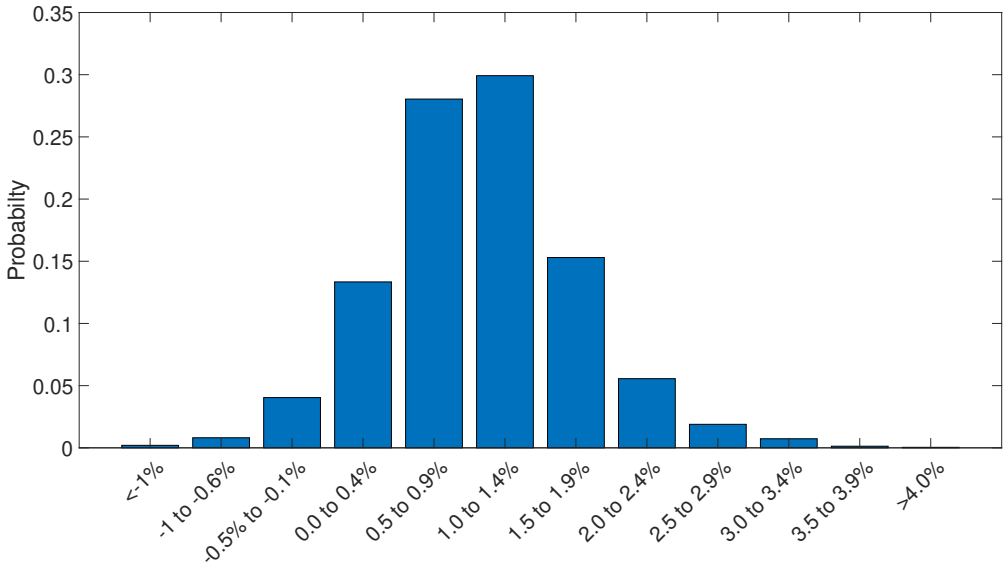
Unemployment				
Survey date	Deadline	Info available	Forecast 1 year	Forecast 2 years
Q1.Y	m1.Y	m11.Y-1	m11.Y	m11.Y+1
Q2.Y	m4.Y	m2.Y	m2.Y+1	m2.Y+2
Q3.Y	m7.Y	m5.Y	m5.Y+1	m5.Y+2
Q4.Y	m10.Y	m8.Y	m8.Y+1	m8.Y+2

Real GDP Growth				
Survey date	Deadline	Info available	Forecast 1 year	Forecast 2 years
Q1.Y	m1.Y	Q3.Y-1	Q3.Y	Q3.Y+1
Q2.Y	m4.Y	Q4.Y-1	Q4.Y	Q4.Y+1
Q3.Y	m7.Y	Q1.Y	Q1.Y+1	Q1.Y+2
Q4.Y	m10.Y	Q2.Y	Q2.Y+1	Q2.Y+2

Note: timing of the ECB SPF survey. For each variable and survey round, the table reports the survey deadline, the latest information available to respondents at the deadline and the rolling horizon forecasts requested. m, Q and Y refer to the month, quarter and year considered.

has taken place in the first month of the quarter, i.e. in January, April, July and October. The questionnaire is sent to the panelists just after the release of the HICP. Participants are asked to provide a forecast for the current calendar year, the following calendar year, the calendar year after that, a long term horizon, a rolling horizon one year ahead of the latest available data and a rolling horizon two years ahead of the latest available data. Table 1 shows timings, information available to forecasters and rolling horizon forecasts requested. For instance, in the 2016.Q1 survey, the latest information available for inflation and unemployment is for December 2015 and for November 2015, respectively. Panelists are required to submit by the second half of January 2016 one-

Figure 1: ECB SPF density forecast for HICP one-year rolling horizon, December 2016



Note: the histogram reports one-year ahead rolling horizon aggregate density forecast for HICP from the 2016.Q1 survey round. Participants are asked to report a probability for the realisation in December 2016 to fall in each bin.

year and two-years ahead rolling horizon forecasts for December 2016 and December 2017 for inflation, and for November 2016 and November 2017 for unemployment. For real GDP growth, the latest information available is 2015.Q3 and one-year and two-years ahead rolling horizon forecasts are for 2016.Q3 and 2017.Q3. For more information on the ECB SPF see Garcia (2003) and Bowles, Friz, Genre, Kenny, Meyler and Rautanen (2007). As an exception to the timing in Table 1 in the case of inflation, at the survey deadline 2007.Q1, the latest realisation available to forecasters was January 2007 instead of December 2006.

To report their density forecasts, participants are given a set of specific ranges and are asked to predict the probability that the target variable will fall in each specific range, or bin, with the first and the last being open intervals. The number of ranges given in every survey round can change but their width is fixed. The ECB SPF reports both the anonymised individual density forecasts and the aggregate density forecast, constructed by summing up the individual probabilities reported in the SPF and dividing by the

number of respondents. For example, Figure 1 reports the aggregate density forecast one year ahead for the HICP for December 2016 produced in the 2016.Q1 survey round.

3 Revisions in Macroeconomic Variables

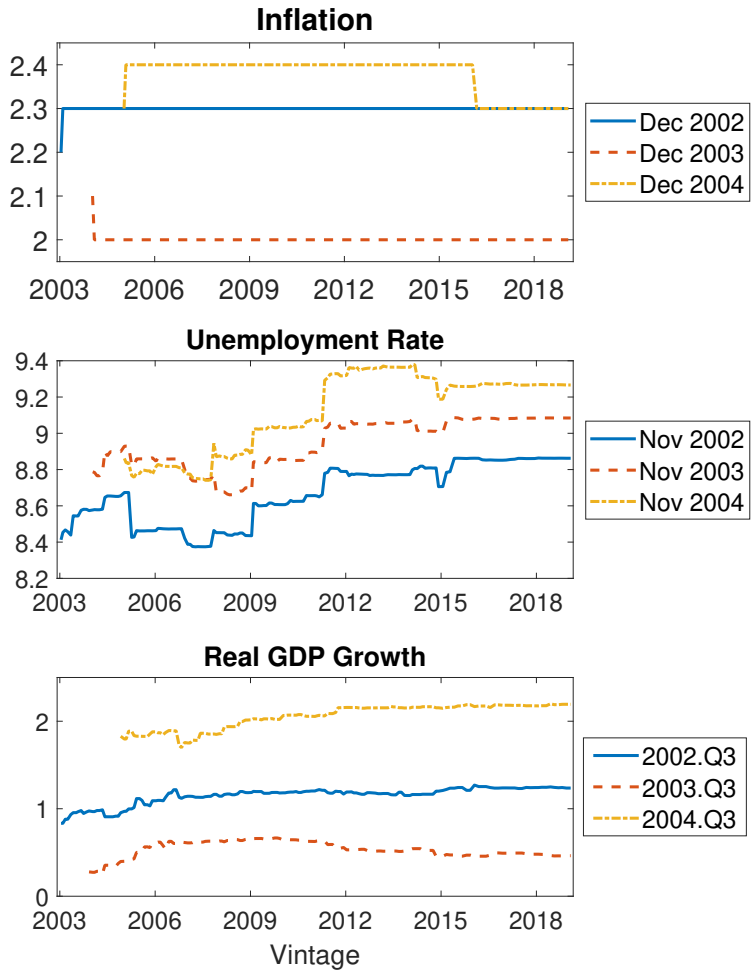
Revisions in macro variables may be caused by new data available, changes in definitions and classifications or correction of clerical mistakes. Mankiw and Shapiro (1986) argue that revisions are most likely caused by unforecastable new information not known at the time data was released.

Figure 2 shows the effect of revisions on macroeconomic data. For instance, the HICP annual growth rate for December 2002 was initially released on 15/01/2003 at 2.2 and, after revisions, it was amended and kept to 2.3 until 30/01/2019. For unemployment in November 2002, the first release was 8.41 on 15/01/2003 and the last release available is 8.86. For real GDP growth, the first release of the third quarter of 2002 was 0.83, but the latest release is 1.24. Overall, the figure indicates that the effect of revision is quite important for unemployment and real GDP growth but minor for inflation.

As forecasters operate using data as available at the time the forecasts are made, and this may be very different from revised data, it is important that revisions in historical realisations of the target variables are accounted for. As shown in Stark (2010), different choices of the vintage of the target variables can impact forecast evaluation. For this reason, we use multiple vintages for the realised values. In particular, we assess forecasting performance using three different releases of the target variable: the first, the fifth and the current release (available on 31/01/2019). Realised values are taken from the Real-time Database for the euro area built by Giannone, Henry, Lalik and Modugno (2012) and available on the European Central Bank Statistical Data Warehouse.

Revisions should be also taken into account when constructing simple benchmark density forecasts to compare with ECB SPF density forecasts. Accordingly, we construct benchmark forecasts using the same vintage of the data available to forecasters when

Figure 2: Revisions in historical data



Note: the three plots show the effect of revision in the realised series of HICP inflation (top plot), unemployment rate (middle plot) and real GDP growth (bottom plot) from the first release to the latest available vintage on the 30/01/2019. The horizontal axis denotes the vintage of the data.

they had to submit their forecast.

4 Benchmark Forecasts

We compare ECB SPF density forecasts against three simple benchmarks: a uniform density forecast that represents an agnostic reply to the survey, a Gaussian random walk density forecast that represents a standard benchmark for forecasting, and a naive forecast based on the lagged ECB SPF density forecast that, as such, incorporates all

the information available at the previous survey round.

Each h-step ahead density forecast i specifies the probability that the variable of interest y_t falls in bin k given the information available at time $t-h$, $f_t^{k,i} = P_{t-h}^i(y_t \in k)$ for $k = 1, \ldots, K$. We construct all the benchmark density forecasts using only the real-time information available to professional forecasters up to the deadline for responding to each survey round.

4.1 Uniform

The uniform benchmark is based on a uniform distribution, with constant probability between two thresholds identified using the historical realisations of the target variable. The two thresholds are the maximum and the minimum of the target variable observed at the latest vintage available at each survey deadline. The constant probability assigned between the thresholds is given by 1 divided by the number of bins of the considered survey round which fall between the thresholds: denoting $a_{t-h} = \min(y_{t-h}, y_{t-h-1}, y_{t-h-2}, \ldots, y_1)$, $b_{t-h} = \max(y_{t-h}, y_{t-h-1}, y_{t-h-2}, \ldots, y_1)$, and n_{t-h} as the number of bins between a_{t-h} and b_{t-h},

$$f_t^{k,U} = \begin{cases} \dfrac{1}{n_{t-h}} & a_{t-h} \leq k \leq b_{t-h} \\ 0 & otherwise \end{cases} \tag{1}$$

With this benchmark, the one-year ahead and the two-years ahead forecasts coincide.

4.2 Gaussian Random Walk

Forecasts for the Gaussian random walk benchmark are drawn from a normal distribution with parameters calculated using the last 20 observations at quarterly frequency at the vintage available at each survey deadline. Under the assumption that the target variable follows a random walk without drift, the conditional expectation is y_{t-h}, so we use

$N(\mu_t, \sigma_t^2)$ with $\mu_t = y_{t-h}$ and $\sigma_t^2 = h \times (19)^{-1} \sum_{i=t-19}^{t} (y_{i-h} - y_{i-h-1})^2$ where h is the forecast horizon in quarters. From this normal distribution, we compute the probability that the realization of the target variable falls inside each bin.

4.3 Naive

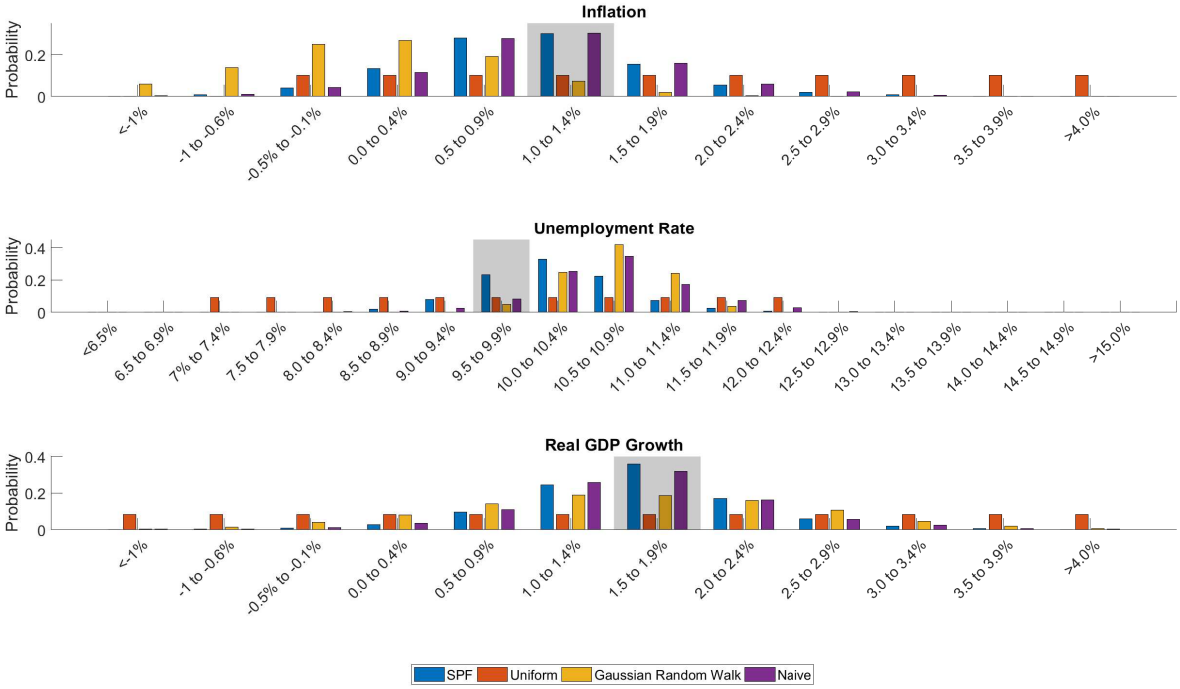
For the naive benchmark, we simply use the last available ECB SPF density forecast for the same horizon, i.e. $f_t^{k,Naive} = f_{t-1}^{k,SPF} = P_{t-h-1}^{SPF}(y_{t-1} \in k)$. In the case of different bins available from a survey round to the following, the forecasts are adjusted to accommodate the new bin structure. If in the new survey round there are more bins than in the previous, the probability of the last bin is equally split across the additional bins available; if there are less bins in the current survey round than the previous round, the probabilities of extreme bins are added up and placed in the only available bin. For additional discussion about the changing bin structure see D'Amico et al. (2008) and Manzan (2017).

We illustrate these three benchmarks with an example. In Figure 3, we report one-year ahead forecasts made at 2016.Q1 for the three benchmarks and the ECB SPF: the blue histograms denote the ECB SPF forecasts, the red histograms denote the uniform benchmark forecasts, the yellow histograms refer to the Gaussian random walk forecasts and the purple histograms denote the naive benchmark forecasts. The shaded areas indicate the interval selected by the realised outcome at the current release. As shown in the figure, the ECB SPF and the benchmarks can have quite different probability distributions over the possible outcomes of the target variable.

5 Density Forecast Evaluation

We compare two h-step ahead density forecasts made at time $t-h$ for the variable of interest y_t using loss functions.

Figure 3: Competing one-year ahead forecasts for the 2016.Q1 survey round



Note: competing one-year ahead forecasts for the 2016.Q1 survey round for inflation (top plot), unemployment rate (middle plot) and real GDP growth (bottom plot). The blue histograms are the ECB SPF forecasts, the red histograms denote the uniform benchmark forecasts, the yellow histograms refer to the Gaussian random walk forecasts and the purple histograms denote the naive benchmark forecasts. The shaded areas indicate the intervals selected by the realised outcomes at the current release.

As shown in the previous section, ECB SPF density forecasts are all reported as histograms, as forecasters report probabilities for intervals, or bins. Two loss functions that naturally accommodate this format of data are the Quadratic Probability Score by Brier (1950) and the Ranked Probability Score by Epstein (1969).

The Quadratic Probability Score (QPS) associated with each forecast is given by

$$QPS_t^i = \sum_{k=1}^{K} (f_t^{k,i} - x_t^k)^2 \tag{2}$$

where the indicator variable $x_t^k = I(y_t \in k)$ takes the value of 1 if the period t outcome falls in bin k and zero otherwise. The QPS loss function penalizes equally any probability assigned to events that do not occur. With this loss function, forecasts that assign a

large probability in a neighborhood of the realised outcome are treated in the same way as forecasts that assign a small probability to that same neighborhood and put more probability on very distant outcomes. This may be appropriate in some situations; in many cases, however, it is desirable to consider more precise the forecast clustering more probability in the intervals near the realised outcome.

The Ranked Probability Score (RPS) associated with each forecast is given by

$$RPS_t^i = \sum_{k=1}^{K} (F_t^{k,i} - X_t^k)^2, \tag{3}$$

where $F_t^{k,i}$ and X_t^k are the cumulative distribution functions of each density forecast $F_t^{k,i} = \sum_{l=1}^{k} f_t^{l,i}$ and of the binary variable $X_t^k = \sum_{l=1}^{k} x_t^k$. The RPS has the advantage of considering the overall tendency of the forecast probability density function, as such it penalizes less severely density forecasts assigning relatively larger probabilities to outcomes that are close to the true outcome. According to Gneiting and Raftery (2007), the RPS has the desirable property of being proper in the sense that encourages the forecasters to reveal their true beliefs.

5.1 Diebold-Mariano-West Test

To evaluate the predictive performance of two density forecasts, we use the Diebold-Mariano-West (DMW) test for the null of equal forecast accuracy by Diebold and Mariano (1995) and West (1996).

Denote by L^i the loss function for $i = 1, 2$, so that $L_t^i = QPS_t^i$ if the QPS loss is used or $L_t^i = RPS_t^i$ if the RPS loss is used, and the loss differential by $d_t = L_t^1 - L_t^2$, the null hypothesis of equal forecasting ability is

$$H_0 : \{E(d_t) = 0\}.$$

Denoting now the sample average as $\overline{d} = \frac{1}{T} \sum_{t=1}^{T} d_t$ and the long run variance as $\sigma_T^2 =$

$var(\sqrt{T}\,\overline{d})$, then the test statistic is $\sqrt{T}\,\overline{d}/\sigma_T$. Under regularity conditions amounting to the existence of a central limit theorem for $\sqrt{T}\,\overline{d}$ and $\sigma_T > 0$, and H_0, then

$$\sqrt{T}\frac{\overline{d}}{\sigma_T} \to_d N(0,1). \tag{4}$$

The test statistic in (4) is unfeasible as σ_T is unknown, but this may be replaced by an estimate, say $\widehat{\sigma}^2$: if the latter is consistent, $\widehat{\sigma} - \sigma_T = o_p(1)$, the feasible statistic obtained in this way retains the standard normal limiting distribution.

One estimate that, under regularity conditions, fits this purpose, is the Weighted Covariance Estimate

$$\widehat{\sigma}^2_{WCE} = \widehat{\gamma}_0 + 2\sum_{j=1}^{T-1} k(j/M)\widehat{\gamma}_j$$

where $\widehat{\gamma}_j = \frac{1}{T}\sum_{t=1}^{T-j}(d_t - \overline{d})(d_{t+j} - \overline{d})$ is the sample autocovariance, and $k(.)$ is a kernel function and M is a bandwidth parameter. Popular kernels include the rectangular kernel

$$k^R(j/M) = 1 \quad \text{for} \quad j \leq M$$

and the triangular (Bartlett) kernel

$$k^B(j/M) = 1 - \frac{j}{M} \quad \text{for} \quad j \leq M$$

yielding the rectangular and the Bartlett estimates, respectively,

$$\widehat{\sigma}^2_{WCE-R} = \widehat{\gamma}_0 + 2\sum_{j=1}^{M} \widehat{\gamma}_j, \qquad \widehat{\sigma}^2_{WCE-B} = \widehat{\gamma}_0 + 2\sum_{j=1}^{M} \left(\frac{M-j}{M}\right)\widehat{\gamma}_j.$$

Regularity conditions to ensure consistency include $M \to \infty$ and $M/T \to 0$ as $T \to \infty$, although Diebold and Mariano (1995) notice that $\widehat{\sigma}^2_{WCE-R}$ is also consistent under the alternative assumption that M is fixed, as long as d_t is a MA(M) process. On the other hand, $\widehat{\sigma}^2_{WCE-R}$ may generate negative (or 0) estimates in a finite sample, a shortcoming

that is not desirable for an estimate of a variance, and may distort the properties of the test.

A second class of estimates of the long run variance is the Weighted Periodogram Estimate

$$\widehat{\sigma}^2_{WPE} = 2\pi \sum_{j=1}^{T/2} K_M(\lambda_j) I(\lambda_j) \tag{5}$$

where $K_M(\lambda_j)$ is a symmetric kernel function, and $I(\lambda_j) = |\frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} d_t e^{i\lambda_j t}|^2$ is the periodogram of d_t evaluated at the Fourier frequencies $\lambda_j = \frac{2\pi j}{T}$ for $j = 1, \ldots, T/2$. A popular kernel in this case is the Daniell kernel

$$K_M^D(j) = \frac{1}{m} \quad \text{for} \quad j \leq m$$

where m is a user chosen parameter that is linked to the bandwidth M (and it is, with slight abuse of notation, referred to as bandwidth too). This kernel is often a convenient choice, as the Daniell kernel estimate of the long run variance has a very simple formula in the frequency domain,

$$\widehat{\sigma}^2_{WPE-D} = 2\pi \frac{1}{m} \sum_{j=1}^{m} I(\lambda_j) \tag{6}$$

When $m \to \infty$ and $m/T \to 0$ as $T \to \infty$, the estimate is consistent. An extension of this class of estimators is introduced in Phillips (2005), that shows that σ_T can be consistently estimated by regressing the series of interest on an orthonormal series (although Phillips (2005) actually only considers a constant long term variance, his argument also applies to the more general context we consider here). This orthonormal series may be a set of trigonometric polynomials but it does not necessarily have to be the case.

Unfortunately, the DMW test is subject to severe size distortion in small and medium sized samples, as documented, for example, in Clark (1999). Obviously, finite sample size distortion is not a problem affecting only the DMW test: it is common to any test that makes inference on the mean (or on a regression parameter) using a heteroskedasticity autocorrelation consistent estimate of the long run variance and maintaining the limit

normality assumption for the standardised statistic, see for example Newey and West (1994). In fact, in any finite sample, the ratio M/T is still non-zero, and in a moderate size sample this ratio may be non-negligible. Thus, this size distortion may be more severe in the context of forecast evaluation, as in many cases the sample is relatively small, when compared to samples in other macro and financial applications.

5.2 Diebold-Mariano-West Test with Fixed-smoothing Asymptotics

Neave (1970) shows that treating the ratio M/T as constant can provide a better measure of the variance of the weighted covariance estimate of a spectral estimate. Kiefer and Vogelsang (2002a,b, 2005) apply the same intuition to the problem of testing hypothesis about the mean for a weakly dependent process, deriving the distribution of the feasible test statistic when $M/T \to b \in (0,1]$ as $T \to \infty$. Under this assumption $\widehat{\sigma}^2$ is not consistent, and the test statistic has a non standard limit distribution that depends on both b and the kernel choice. Because of the dependence on b of the limit distribution, this approach is often referred to as "fixed-b".

In the context of the DMW test, for the Bartlett kernel the results of Kiefer and Vogelsang (2005) imply that, under H_0 and regularity conditions, when $M/T \to b \in (0,1]$ as $T \to \infty$,

$$\sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_{WCE-B}} \to_d \Phi^B(b) \tag{7}$$

$\Phi^B(b)$ is characterised in Kiefer and Vogelsang (2005) and a cubic equation is provided for critical values.

In the frequency domain, fixed-b corresponds to keeping m constant when the Daniell kernel is used. This naturally leads to considering asymptotics for fixed m. Under H_0 and regularity conditions, Hualde and Iacone (2017) consider m constant as $T \to \infty$, in

this case we have

$$\sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_{WPE-D}} \rightarrow_d t_{2m}. \tag{8}$$

Sun (2013) shows that a limit of this kind also holds for the general orthonormal series variance estimator.

Fixed-b and fixed-m asymptotics can be heuristically understood as undersmoothing in the context of estimating the spectral density at frequency zero: for this reason, many references, for example Sun (2013), refers to them collectively as fixed-smoothing.

Monte Carlo simulations in Kiefer and Vogelsang (2005) suggest that critical values obtained using fixed-b asymptotics result in better empirical size for tests. This was later justified theoretically by Sun (2014), that shows that fixed-b asymptotics provides a higher order refinement. Moreover, fixed-smoothing asymptotics gives a justification (and suitable critical values) even for bandwidths that researchers would not consider when using standard asymptotics: it is even possible to choose $M = T$ when using the weighted covariance Bartlett estimate, or to choose $m = 1$ when using the weighted periodogram Daniell estimate. This allows a further correction in the empirical size, as in Monte Carlo simulations larger bandwidths M (smaller m) are associated to better empirical size: for example, Monte Carlo simulations in Coroneo and Iacone (2015) indicate that it is possible to completely eliminate the size distortion documented by Clark (1999).

Assumption 1 *Partial sums of d_t are such that the functional central limit theorem (FCLT) holds*

$$\frac{\sqrt{T}}{T}\frac{1}{\sigma_T}\sum_{t=1}^{\lfloor rT \rfloor} d_t \Rightarrow W(r)$$

where $\lfloor . \rfloor$ denotes the integer part of a number, $r \in [0,1]$ and $W(r)$ is a standard Brownian motion.

Assumption 1 is sufficient to establish the fixed smoothing limits (7) and (8). This assumption is not primitive, but it is convenient because it may be established under a

range of conditions. For example, Phillips and Solo (1992) consider linear processes of independent, identically distributed innovations of martingale difference sequences. On the other hand, Wooldridge and White (1988) consider mixing processes, thus allowing for forms of heteroskedasticity that may also induce non-stationarity, under the additional assumption that $Var\left(\sqrt{T}/T\ 1/\sigma_T \sum_{t=1}^{\lfloor rT \rfloor} d_t\right) \to r$. In view of the non-linearity in the loss function, establishing a linear representation for d_t from primitive assumptions on $f_t^{k,i}$ and x_t^k may be very challenging, whereas establishing mixing properties may be easier, especially when $f_t^{k,i}$ and x_t^k are limited to being M-dependent processes. However, as the two classes may overlap but are not included in each other, see discussion in Phillips and Solo (1992) and Andrews (1984), we prefer the more general assumption given here, that encompasses them both.

6 Monte Carlo Study of Size and Power

We analyse the empirical size and power by means of a Monte Carlo experiment. Since Kiefer and Vogelsang (2005), simulation studies have by now covered a fairly wide range of situations, including inference in regression models, in non-linear models, and others: we refer to Lazarus, Lewis, Stock and Watson (2018) for a recent, comprehensive study. In point forecasting, studies include Coroneo and Iacone (2015) on forecast evaluation in small samples, Harvey, Leybourne and Whitehouse (2017) on forecast encompassing, and Li and Patton (2018) on forecast evaluation in large samples.

We already noticed that simulation studies find that fixed-smoothing asymptotics yield better approximation of the empirical size, and that this improvement is stronger the larger is the bandwidth M (the smaller is m). These works also find that the finite sample power is decreasing with the bandwidth, therefore documenting the existence of a trade-off between correct size and power: Lazarus, Lewis, Stock and Watson (2018), drawing on their extensive simulation study, recommend $M = \lfloor 1.3T^{1/2} \rfloor$ and $m = \lfloor 0.2T^{2/3} \rfloor$.

In this section we check whether the size improvements for the DMW test still hold in the case of QPS and RPS loss functions, and in a much smaller sample, that replicates the dimension of the sample of our dataset. We also focus our analysis on the issue of bandwidth selection and compare our results with Lazarus, Lewis, Stock and Watson (2018) and Coroneo and Iacone (2015).

In our Monte Carlo study, we consider a sample of T observations, and we assume that y_t may take any integer value between 1 and κ, where κ is odd and K is such that $\kappa \leq K$. The probability distribution of y_t depends on t and y_{t-1}, as follows:

- For $t = 1, (Q+1)+1, 2(Q+1)+1, 3(Q+1)+1, \ldots$, where Q is a fixed integer and such that $Q < T$, we assume for the data generating process that $P(y_t = k) = 1/\kappa$ for any $k = 1, \ldots \kappa$.

- For other values of t, the probabilities are assigned according to the previous realisation of y_t. If we define $\pi = \lfloor \kappa/2 \rfloor + 1$, we have

  - $P(y_t = k | y_{t-1} < \pi) = 1/\pi$ for $k = 1, \ldots \pi$, and 0 otherwise;
  - $P(y_t = k | y_{t-1} = \pi) = 1/\kappa$ for $k = 1, \ldots \kappa$, and 0 otherwise;
  - $P(y_t = k | y_{t-1} > \pi) = 1/\pi$ for $k = \pi, \ldots \kappa$, and 0 otherwise.

Notice that after Q periods the data generating process (DGP) resets and y_t is then taken again from a discrete uniform distribution spanning $(1, \kappa)$.

We set the forecasting rule 1 as $f_t^{k,1} = P^1(y_t = k) = 1/\pi$ when $k = 1 \ldots \pi$, and 0 otherwise; and the forecasting rule 2 as $f_t^{k,2} = P^2(y_t = k) = 1/\pi$ when $k = \pi + c \ldots 2\pi - 1 + c$, and 0 otherwise, the parameter c being an integer and such that $\kappa + c = K$.

Example 1 *With $\kappa = 5$ and $Q = 1$, the variable of interest is generated by the following DGP:*

- *for $t = 1$, $P(y_1 = k) = 1/5$ for $k = 1, \ldots, 5$,*

- *for $t = 2$*

- $P\left(y_2 = k | y_1 = 1 \text{ or } 2\right) = 1/3$ *for* $k = 1, \ldots, 3$*;*
- $P\left(y_2 = k | y_1 = 3\right) = 1/5$ *for* $k = 1, \ldots, 5$*;*
- $P\left(y_2 = k | y_1 = 4 \text{ or } 5\right) = 1/3$ *for* $k = 3, \ldots, 5$*.*

- *for* $t = 3$, $P\left(y_3 = k\right) = 1/5$ *for* $k = 1, \ldots, 5$*,*

and so on. With $c = 0$ *the forecasting rules are* $f_t^{k,1} = 1/3$ *when* $k = 1 \ldots 3$ *and* $f_t^{k,2} = 1/3$ *when* $k = 3 \ldots 5$*.*

Clearly, with this data generating process, when $c = 0$ *it holds that* $E(d_t) = 0$*. Notice that this process is not stationary, because for example* $Var(d_1) \neq Var(d_2)$ *and* $cov(d_1, d_2) \neq 0$ *but* $cov(d_2, d_3) = 0$*. However, the process* d_t *is by construction* M*-dependent, and the regularity condition on the variance of the partial sums also holds, so the FCLT still holds here, see Appendix A for a more intuitive argument.*

When $c \neq 0$ the null hypothesis does not hold, as the competing forecast $f_t^{k,2}$ is shifted by c and, given our DGP, on average this should be less precise than $f_t^{k,1}$, so we can investigate the power as we increase the value of c. When $Q \geq 1$ the process is dependent, and the dependence increases with Q and also with κ.

In our experiment we set $\kappa = 21$ and $\kappa = 5$ and Q up to 4, with sample size set at $T = 30$, 60, and we repeat the experiment for 10,000 replications. Notice that our sample size is much smaller than the sample size of Lazarus, Lewis, Stock and Watson (2018), and it matches the dimension of the sample available for our empirical study. Indeed, checking the empirical performance in such small samples is one reason of interest in this experiment.

In Tables 2-5 we report the empirical size of the test when critical values from both standard asymptotics and fixed-smoothing asymptotics are used. In columns WCE, the long run variance estimate is computed using a Bartlett kernel with bandwidths $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$ and $M = T$, except for the column DM for which we use the rectangular kernel with $M = Q$ as suggested by Diebold and Mariano (1995). In

Table 2: Empirical size of the DMW test with standard asymptotics, $T = 60$

Ranked Probability Score

$\kappa = 21$

		WCE			WPE			
Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.073	0.067	0.084	0.331	0.117	0.094	0.071	0.061
1	0.088	0.095	0.104	0.340	0.125	0.103	0.083	0.088
2	0.096	0.116	0.113	0.350	0.125	0.099	0.088	0.131
3	0.112	0.156	0.127	0.352	0.128	0.110	0.104	0.190
4	0.124	0.183	0.136	0.346	0.130	0.111	0.124	0.224

$\kappa = 5$

Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.074	0.067	0.087	0.329	0.119	0.097	0.071	0.060
1	0.084	0.091	0.103	0.341	0.125	0.104	0.079	0.084
2	0.097	0.113	0.110	0.347	0.124	0.098	0.086	0.117
3	0.116	0.135	0.121	0.346	0.133	0.110	0.096	0.149
4	0.125	0.139	0.119	0.341	0.123	0.101	0.098	0.159

Quadratic Probability Score

$\kappa = 21$

		WCE			WPE			
Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.073	0.066	0.086	0.333	0.121	0.100	0.070	0.062
1	0.086	0.097	0.106	0.339	0.128	0.105	0.080	0.091
2	0.095	0.121	0.112	0.355	0.123	0.101	0.088	0.135
3	0.112	0.157	0.127	0.353	0.127	0.109	0.105	0.196
4	0.124	0.190	0.139	0.349	0.132	0.112	0.128	0.238

$\kappa = 5$

Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.071	0.066	0.087	0.330	0.122	0.099	0.068	0.058
1	0.085	0.091	0.104	0.340	0.125	0.105	0.080	0.083
2	0.099	0.110	0.108	0.348	0.124	0.101	0.084	0.115
3	0.114	0.137	0.120	0.349	0.129	0.108	0.097	0.152
4	0.124	0.140	0.117	0.345	0.124	0.102	0.099	0.161

Note: the table reports the empirical size of the Diebold-Mariano-West test with standard asymptotics and sample size $T = 60$. The theoretical size is 5%. κ is an odd number indicating the number of possible realisations of the target variable. Q indicates the number of periods after which the DGP resets. The higher the Q, the higher the level of dependence in the process. WCE refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance (except for column DM, where the rectangular kernel is used); WCE refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance.

Table 3: Empirical size of the DMW test with fixed-smoothing asymptotics, $T = 60$

Ranked Probability Score

$\kappa = 21$

	WCE			WPE			
Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.045	0.045	0.049	0 .049	0.053	0.051	0.050
1	0.069	0.058	0.057	0 .054	0.056	0.058	0.076
2	0.090	0.061	0.059	0 .054	0.051	0.063	0.116
3	0.122	0.074	0.065	0 .058	0.058	0.080	0.173
4	0.149	0.083	0.070	0 .058	0.059	0.099	0.208

$\kappa = 5$

Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.045	0.044	0.048	0.050	0.052	0.050	0.052
1	0.064	0.056	0.057	0.053	0.054	0.058	0.073
2	0.082	0.059	0.056	0.051	0.052	0.062	0.106
3	0.102	0.070	0.062	0.054	0.057	0.071	0.134
4	0.108	0.068	0.062	0.052	0.058	0.076	0.143

Quadratic Probability Score

$\kappa = 21$

	WCE			WPE			
Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.045	0.045	0.050	0.050	0.054	0.051	0.052
1	0.071	0.056	0.057	0.055	0.058	0.059	0.080
2	0.093	0.064	0.058	0.053	0.052	0.065	0.123
3	0.124	0.076	0.065	0.055	0.058	0.079	0.178
4	0.155	0.085	0.073	0.060	0.062	0.101	0.217

$\kappa = 5$

Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.043	0.043	0.047	0.052	0.053	0.050	0.049
1	0.065	0.055	0.057	0.053	0.057	0.056	0.073
2	0.083	0.061	0.057	0.052	0.053	0.063	0.104
3	0.101	0.070	0.059	0.054	0.058	0.073	0.137
4	0.109	0.069	0.062	0.053	0.057	0.075	0.145

Note: the table reports the empirical size of the Diebold-Mariano-West test with fixed smoothing asymptotics and sample size $T = 60$. The theoretical size is 5%. κ is an odd number indicating the number of possible realisations of the target variable. Q indicates the number of periods after which the DGP resets. The higher the Q, the higher the level of dependence in the process. WCE refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance; WCE refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance.

Table 4: Empirical size of the DMW test with standard asymptotics, $T = 30$

Ranked Probability Score

$\kappa = 21$

		WCE			WPE			
Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.098	0.080	0.102	0.335	0.119	0.099	0.077	0.062
1	0.123	0.114	0.126	0.348	0.132	0.106	0.091	0.110
2	0.157	0.149	0.150	0.357	0.138	0.116	0.120	0.181
3	0.202	0.186	0.171	0.378	0.146	0.134	0.158	0.236
4	0.210	0.222	0.197	0.367	0.148	0.157	0.201	0.269

$\kappa = 5$

Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.100	0.083	0.104	0.335	0.120	0.099	0.079	0.064
1	0.126	0.112	0.124	0.351	0.127	0.107	0.091	0.107
2	0.160	0.138	0.142	0.355	0.133	0.115	0.111	0.156
3	0.204	0.160	0.156	0.361	0.138	0.124	0.132	0.186
4	0.229	0.177	0.167	0.362	0.137	0.130	0.152	0.210

Quadratic Probability Score

$\kappa = 21$

Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.102	0.083	0.106	0.326	0.124	0.100	0.079	0.065
1	0.128	0.117	0.130	0.346	0.131	0.106	0.096	0.119
2	0.159	0.147	0.151	0.356	0.137	0.118	0.123	0.182
3	0.205	0.188	0.169	0.374	0.143	0.131	0.156	0.256
4	0.211	0.230	0.203	0.371	0.161	0.169	0.204	0.276

$\kappa = 5$

		WCE			WPE			
Q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.101	0.083	0.105	0.339	0.123	0.100	0.078	0.064
1	0.130	0.111	0.126	0.350	0.128	0.107	0.092	0.107
2	0.160	0.139	0.143	0.357	0.134	0.115	0.112	0.158
3	0.205	0.162	0.157	0.367	0.136	0.124	0.134	0.188
4	0.227	0.180	0.167	0.361	0.136	0.132	0.151	0.213

Note: the table reports the empirical size of the Diebold-Mariano-West test with standard asymptotics and sample size $T = 30$. The theoretical size is 5%. κ is an odd number indicating the number of possible realisations of the target variable. Q indicates the number of periods after which the DGP resets. The higher the Q, the higher the level of dependence in the process. WCE refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance (except for column DM, where the rectangular kernel is used); WCE refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance.

Table 5: Empirical size of the DMW test with fixed-smoothing asymptotics, $T = 30$

Ranked Probability Score

$\kappa = 21$

	WCE			WPE			
Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.040	0.040	0.045	0.047	0.049	0.051	0.049
1	0.066	0.060	0.058	0.053	0.058	0.063	0.091
2	0.094	0.075	0.071	0.060	0.066	0.087	0.154
3	0.128	0.096	0.080	0.059	0.081	0.121	0.208
4	0.166	0.112	0.084	0.058	0.089	0.164	0.243

$\kappa = 5$

Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.042	0.041	0.046	0.051	0.052	0.052	0.050
1	0.062	0.057	0.056	0.054	0.058	0.061	0.086
2	0.082	0.069	0.065	0.057	0.060	0.076	0.130
3	0.102	0.079	0.075	0.061	0.071	0.097	0.157
4	0.117	0.090	0.079	0.063	0.077	0.113	0.183

Quadratic Probability Score

$\kappa = 21$

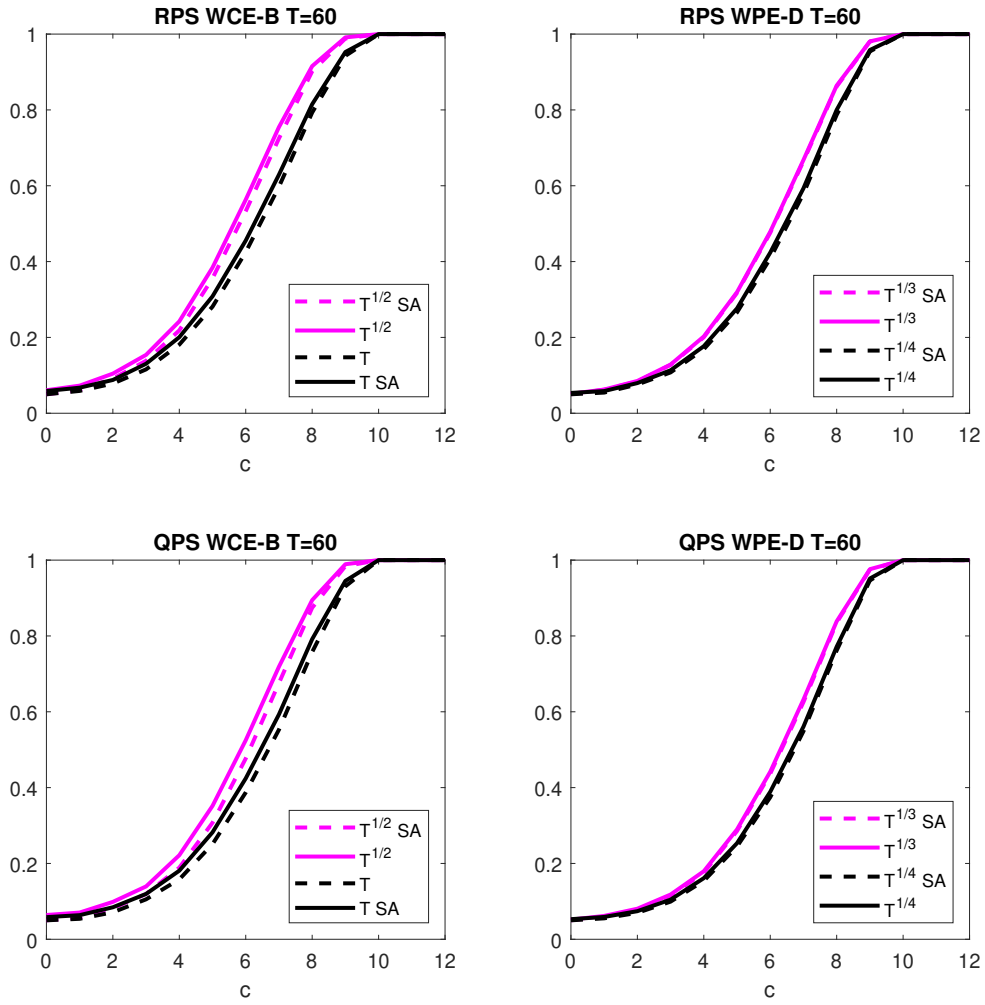
	WCE			WPE			
Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.041	0.040	0.048	0.050	0.050	0.051	0.051
1	0.070	0.063	0.057	0.055	0.061	0.064	0.101
2	0.104	0.082	0.075	0.064	0.066	0.097	0.157
3	0.131	0.099	0.084	0.060	0.086	0.125	0.226
4	0.179	0.121	0.084	0.053	0.075	0.175	0.254

$\kappa = 5$

Q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
0	0.042	0.042	0.047	0.053	0.051	0.049	0.049
1	0.062	0.058	0.056	0.055	0.058	0.062	0.085
2	0.083	0.069	0.068	0.057	0.060	0.078	0.133
3	0.101	0.079	0.074	0.061	0.070	0.095	0.161
4	0.120	0.090	0.081	0.063	0.077	0.114	0.187

Note: the table reports the empirical size of the Diebold-Mariano-West test with fixed smoothing asymptotics and sample size $T = 30$. The theoretical size is 5%. κ is an odd number indicating the number of possible realisations of the target variable. Q indicates the number of periods after which the DGP resets. The higher the Q, the higher the level of dependence in the process. WCE refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance; WCE refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance.

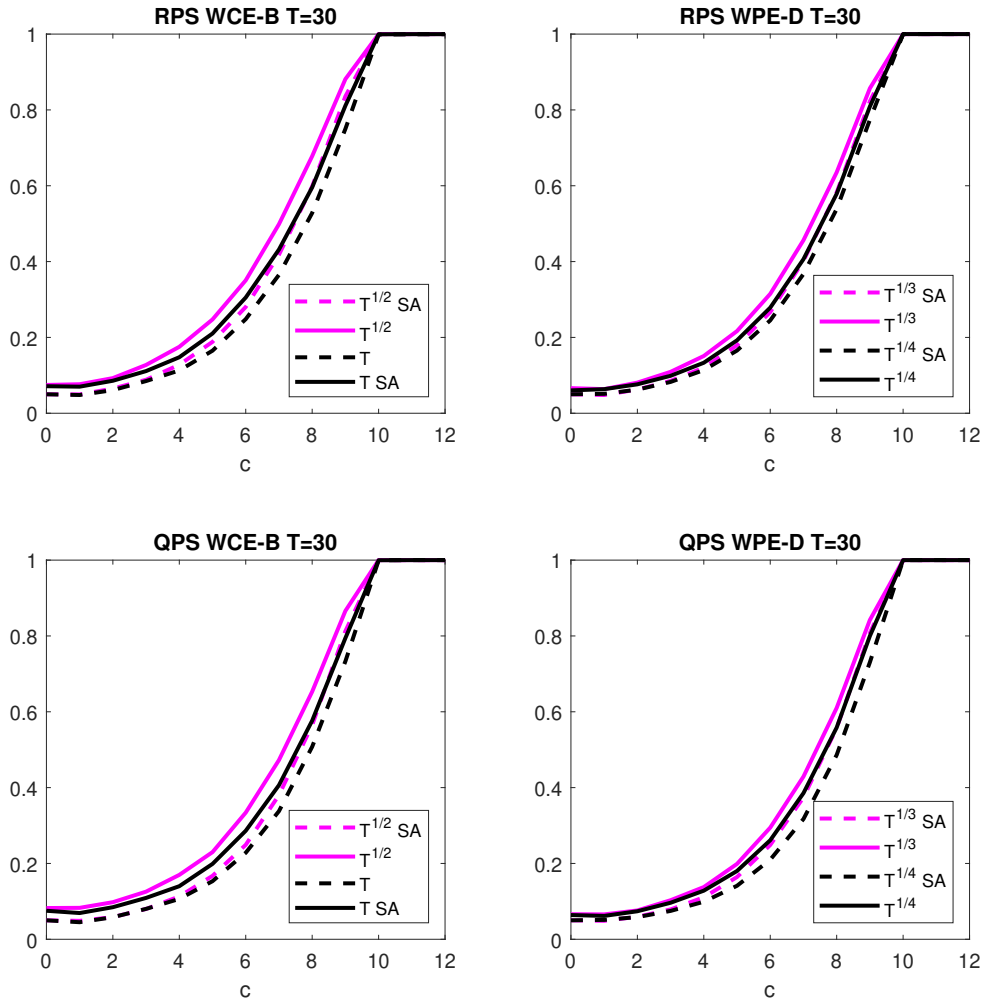
Figure 4: Finite sample local power, $T = 60$.



Note: the figure reports power performances of the Diebold-Mariano-West test with Quadratic Probability Score (QPS) and Ranked Probability Score (RPS) in a sample of size $T = 60$. The dashed lines refer to power performances using size-adjusted critical values while solid lines use fixed-smoothing asymptotics. The parameter c indicates the distance from the null hypothesis. WCE-B refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance; WPE-D refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance. Data generating process with the number of possible realisations of the target variable $\kappa = 21$ and the number of periods after which the DGP resets $Q = 3$.

columns WPE, we use the Daniell kernel with bandwidths $m = \lfloor T^{1/4} \rfloor$, $m = \lfloor T^{1/3} \rfloor$, $m = \lfloor T^{1/2} \rfloor$ and $m = \lfloor T^{2/3} \rfloor$. Consistently with results from other simulation studies, standard asymptotics are associated with size distortion, the performance worsening as the dependence increases with Q. Using fixed smoothing asymptotics always improves

Figure 5: Finite sample local power, $T = 30$.



Note: the figure reports power performances of the Diebold-Mariano-West test with Quadratic Probability Score (QPS) and Ranked Probability Score (RPS) in a sample of size $T = 30$. The dashed lines refer to power performances using size-adjusted critical values while solid lines use fixed-smoothing asymptotics. The parameter c indicates the distance from the null hypothesis. WCE-B refers to the test statistic with Weighted Covariance Estimate with Bartlett kernel for the long run variance; WPE-D refers to the test statistic with Weighted Periodogram Estimate with Daniell kernel for the long run variance. Data generating process with the number of possible realisations of the target variable $\kappa = 21$ and the number of periods after which the DGP resets $Q = 3$.

the size: on balance, we observe correctly sized tests with WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ and, better still, $M = T$; likewise, we observe correct size with WPE and bandwidth $m = \lfloor T^{1/3} \rfloor$ and, better still, $m = \lfloor T^{1/4} \rfloor$. The results do not seem to be sensitive to the loss function, whether QPS or RPS.

For the power study, we set $\kappa = 21$ and $Q = 3$ and increasing values of c. In this case, we only consider bandwidths that are associated to good empirical size properties, namely WCE with $M = \lfloor T^{1/2} \rfloor$ and $M = T$, and WPE with $m = \lfloor T^{1/4} \rfloor$ and $m = \lfloor T^{1/3} \rfloor$, in all cases only for fixed-smoothing asymptotics. For the purpose of comparison only, we also plot the size adjusted power. Power performances are reported in Figures 4 and 5. In all cases the empirical power is a good approximation of the size adjusted power, again offering support to the assumption that fixed-smoothing asymptotic is a valuable instrument for inference, although we appreciate that a small deviation between the feasible and size adjusted power curves still exist, when the WCE estimate is used, especially in the smaller sample and for the shorter bandwidth. We also find that, as a general rule, larger bandwidths M (smaller m) are associated to lower power, consistently with other similar simulation studies. Overall we suggest $M = \lfloor T^{1/2} \rfloor$ and $m = \lfloor T^{1/3} \rfloor$: given our sample size, these bandwidth rules seem in line with Lazarus, Lewis, Stock and Watson (2018). Finally, it is important to notice that the power curves do not seem to differ much, depending on the loss function. There would be no reason, therefore, to prefer the RPS to the QPS on size and power only.

7 Empirical Results

We analyse the SPF rolling horizon aggregate density forecasts as reported by the ECB for HICP inflation, unemployment rate and real GDP growth, for the surveys between 2001.Q1 and 2016.Q2, corresponding to a total of 62 quarterly observations. We do not evaluate the predictive ability of core inflation forecasts due to the unavailability of data in the sample considered.

Special questionnaires were sent in the autumn of 2008, 2013 and 2018 asking participants about their forecasting practices, see de Vincent-Humphreys, Dimitrova, Falck and Henkel (2019) for a detailed discussion. Responses indicate that forecasts are based on one or more models to cross check results but, especially for long term forecasts, judg-

ment plays an important role, with one third of respondents reporting that their forecasts are essentially judgment based. Moreover, the majority of participants reported that the importance of judgment increased following the financial crisis. Therefore we also split the sample in two equally sized sub-samples: 2001.Q1-2008.Q3 and 2008.Q4-2016.Q2, of 31 observations each. As shown in Section 6, with such small sample sizes the DMW test with standard asymptotics suffers from large size distortions but fixed-smoothing asymptotics can still provide reliable inference.

In Figures 6–8, we plot the realised values for each variable (using the current release) along with the median and the 15% and 85% quantiles of the density forecasts of the three benchmarks and the ECB SPF. The bottom plots of the figures report the RPS associated with the ECB SPF, the Gaussian random walk and the naive benchmark.

DMW test statistic values are reported in Tables 6, 9 and 12 for the full sample and in Tables 7, 8, 10, 11, 13 and 14 for the sub-samples. A negative value of the test indicates that the benchmark is performing better than the ECB SPF forecast, while a positive value indicates that the ECB SPF is predicting better than the benchmark. Rejections from standard asymptotics critical values are indicated shading the appropriate cell; and indicate, respectively, two-sided significance at the 5% and 10% level. Rejections using fixed-smoothing asymptotics critical values are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

The top panel of each table uses the Ranked Probability Score, and the bottom panel uses the Quadratic Probability Score. Density forecasts are compared to the three benchmarks: uniform, Gaussian random walk and naive. The long run variance is estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). We take realised values from the real-time database by Giannone, Henry, Lalik and Modugno (2012) and use three different vintages of the realised variable: first release, fifth release (four releases

after the first) and the latest available release on 30/01/2019 (current release).

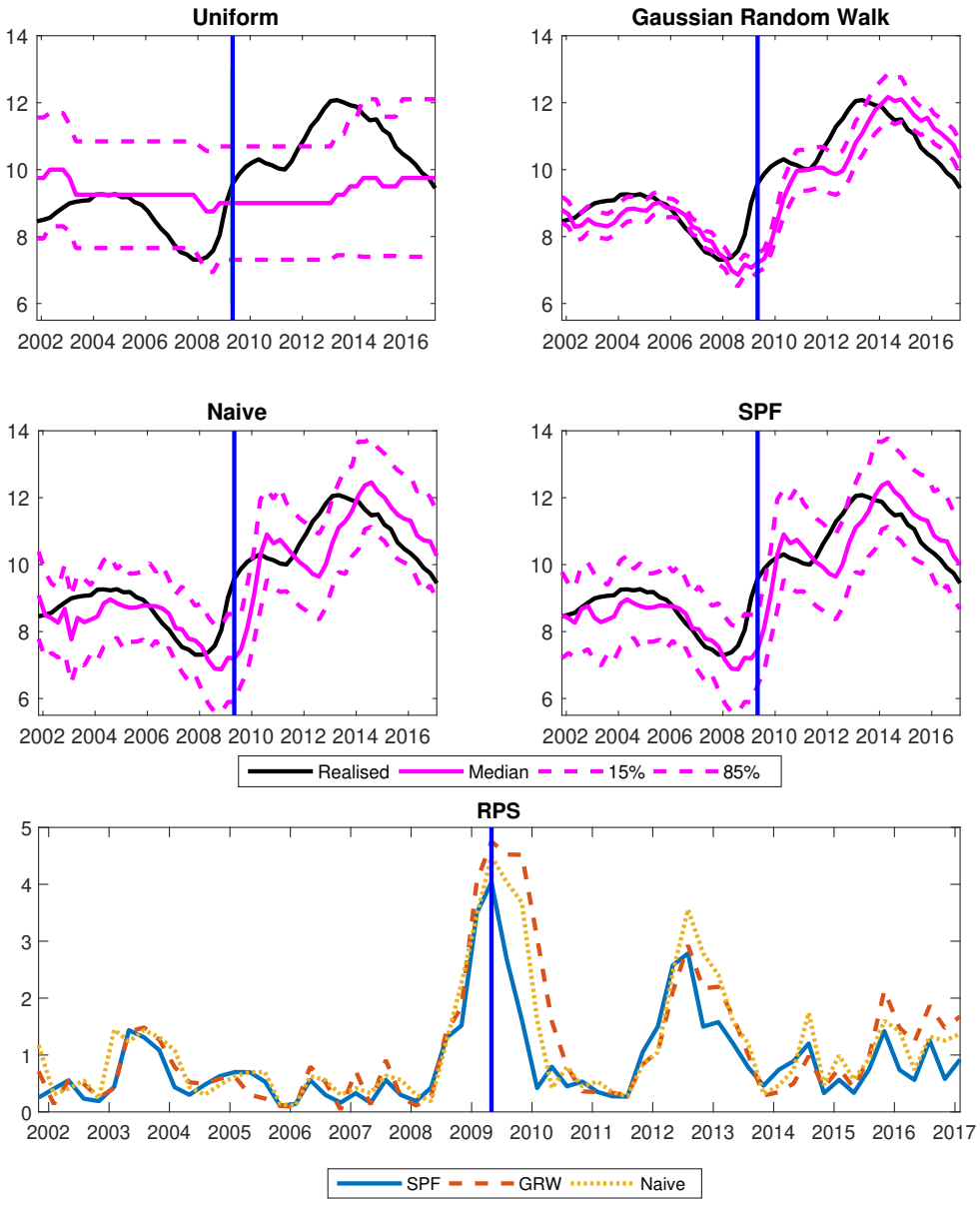
For the unemployment rate, Figure 6 shows that the three benchmarks imply very different density forecasts. The uniform benchmark has the largest dispersion for given date, while the Gaussian random walk predictions are more concentrated for given date. The naive benchmark density forecast is similar to the ECB SPF one, as it is constructed using the previous ECB SPF forecast, and by comparing them we can see the value of the latest information: the bottom panel indicates that the RPS of the naive is nearly always larger than the RPS of the ECB SPF. For the Gaussian random walk, although there are period where the RPS is larger than that of the ECB SPF, there are also periods where the reverse is true. Results in Tables 6–8, indicate that one-year ahead ECB SPF forecasts appear to mostly outperform the uniform and the naive benchmarks in all three samples; the evidence for the Gaussian random walk is weaker, especially in the subsamples. For the two-years ahead forecasts, the surveys also appear to beat the naive benchmark. We can also notice that the ECB SPF forecasts superiority is more marked in the second half of the sample.

Testing real GDP growth gives similar results to those obtained for unemployment, with ECB SPF forecasts superior for one-year ahead predictions, especially in the second half of the sample. However, looking at individual benchmarks we can see that the ECB SPF significantly outperforms the naive benchmark in the full sample and in the first subsample. In the second subsample, instead the ECB SPF significantly outperforms the uniform and the Gaussian random walk.

Taking these results together, we conclude that, for unemployment and the real GDP growth rate, the professional forecasters outperform simple benchmarks at least at one-year horizon, indicating that they use more sophisticated models to extract the available information.

The plots of forecasts and realisations for HICP are in figure 8: we first notice that inflation is close to the target in the first half of the sample and up to the 2008 crisis;

Figure 6: Unemployment rate (one-year rolling horizon forecast)



Note: The four top plots report the realised unemployment rate at the current release (solid black line) along with the median of the forecast distributions (solid pink line), and the 15% and 85% quantiles of the forecast distributions (the pink dotted lines). The bottom plot reports the Ranked Probability Score at the current release for the SPF forecasts (light blue solid line), the Gaussian random walk forecasts (red dashed line) and the naive forecasts (yellow dotted line). Forecasts are for November 2001 to February 2017. The blue vertical line in May 2009 separates the two sub-samples considered.

Table 6: DMW test for the unemployment rate. Full sample Q1.2001 - Q2.2016, $T = 62$.

Ranked Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	4.47	2.17	3.50	0.46	1.79	2.46
WCE-B	3.78**	2.05*	4.16**	0.51	1.76	2.88**
WPE-D	2.98**	2.17*	4.96**	0.51	1.44	4.29**
			Fifth release			
WCE-DM	4.08	2.30	3.40	0.40	1.88	2.34
WCE-B	3.43**	2.17*	4.34**	0.43	1.82	2.79**
WPE-D	2.76**	2.25*	5.38**	0.43	1.46	4.22**
			Current release			
WCE-DM	2.97	2.14	3.24	-0.16	1.58	2.19
WCE-B	2.52**	1.97*	3.68**	-0.17	1.55	2.59**
WPE-D	2.08*	1.84	4.19**	-0.17	1.29	3.76**

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	2.64	2.36	3.34	-0.41	1.88	3.11
WCE-B	2.75**	2.06*	3.95**	-0.52	1.80	3.26**
WPE-D	3.04**	1.94*	4.54**	-0.68	1.43	3.62**
			Fifth release			
WCE-DM	1.93	2.51	3.62	-0.56	1.74	3.20
WCE-B	1.91*	2.25*	4.46**	-0.70	1.66	3.35**
WPE-D	2.23*	1.91*	4.18**	-0.90	1.29	3.54**
			Current release			
WCE-DM	1.69	2.67	3.32	-1.33	1.85	1.88
WCE-B	1.68	2.47**	3.80**	-1.60	1.82	2.22*
WPE-D	1.55	2.40**	3.50**	-1.77	1.57	5.07**

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the unemployment rate against the uniform, the Gaussian random walk and the naive benchmark forecasts on the full sample Q1.2001 - Q2.2016 (T = 62). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Table 7: DMW test for the unemployment rate. Sub-sample Q1.2001 - Q3.2008, $T = 31$.

Ranked Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	3.43	1.78	2.73	0.02	1.16	1.61
WCE-B	3.28**	1.68	2.57**	0.02	1.22	1.92
WPE-D	2.78**	1.54	2.56**	0.01	1.13	1.68
			Fifth release			
WCE-DM	2.93	2.13	3.10	0.03	1.16	1.55
WCE-B	2.80**	2.10*	3.25**	0.03	1.21	1.86
WPE-D	2.41*	1.74	2.68**	0.02	1.11	1.62
			Current release			
WCE-DM	1.20	1.72	3.49	-0.94	0.37	1.06
WCE-B	1.16	1.77	3.74**	-1.02	0.39	1.29
WPE-D	1.03	1.33	3.19**	-0.88	0.37	1.27

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	1.95	1.50	1.39	-0.28	1.12	1.95
WCE-B	1.99	1.38	1.58	-0.32	1.17	2.06*
WPE-D	1.68	1.22	1.50	-0.25	1.11	1.83
			Fifth release			
WCE-DM	1.76	1.96	2.25	-0.17	0.99	2.48
WCE-B	1.72	1.88	2.48**	-0.20	1.03	2.57**
WPE-D	1.39	1.64	2.14*	-0.15	0.95	2.27*
			Current release			
WCE-DM	1.37	1.82	2.20	-1.31	0.89	0.48
WCE-B	1.40	1.74	2.73**	-1.47	0.94	0.55
WPE-D	1.32	1.28	2.68**	-1.19	0.88	0.50

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the unemployment rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the sub-sample Q1.2001 - Q3.2008 (T = 31). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Table 8: DMW test for the unemployment rate. Sub-sample Q4.2008 - Q2.2016, $T = 31$.

Ranked Probability Score

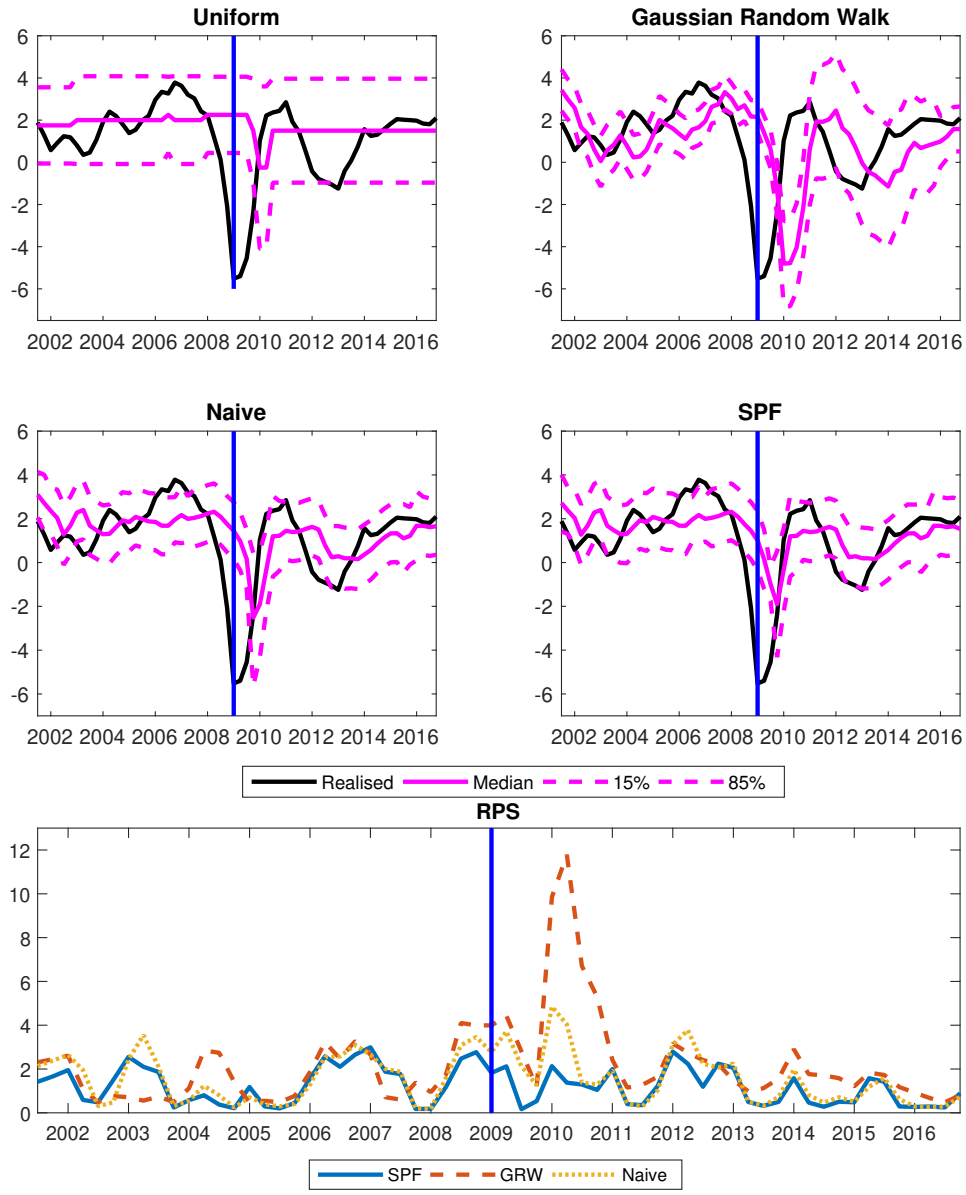
	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	3.33	1.77	2.97	0.71	1.46	2.08
WCE-B	2.93**	1.71	3.67**	0.77	1.53	2.19*
WPE-D	2.53**	1.41	3.24**	0.63	1.35	1.76
			Fifth release			
WCE-DM	3.16	1.77	2.59	0.59	1.57	1.96
WCE-B	2.76**	1.71	3.16**	0.65	1.63	2.09*
WPE-D	2.41*	1.40	2.71**	0.53	1.45	1.68
			Current release			
WCE-DM	3.46	1.90	2.48	0.77	1.67	2.04
WCE-B	3.11**	1.85	2.75**	0.84	1.76	2.18*
WPE-D	2.67**	1.55	2.22*	0.68	1.55	1.75

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	2.39	1.83	3.79	-0.35	1.74	2.72
WCE-B	2.77**	1.69	4.51**	-0.39	1.79	2.51**
WPE-D	2.22*	1.44	4.51**	-0.30	1.59	2.12*
			Fifth release			
WCE-DM	1.16	1.58	3.09	-0.86	1.71	2.02
WCE-B	1.32	1.47	3.51**	-0.97	1.75	1.96
WPE-D	1.07	1.24	3.49**	-0.73	1.54	1.60
			Current release			
WCE-DM	1.19	1.99	2.69	-0.55	1.98	2.96
WCE-B	1.35	1.82	2.75**	-0.62	2.09*	2.87**
WPE-D	1.06	1.56	2.43*	-0.47	1.82	2.30*

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the unemployment rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the sub-sample Q4.2008 - Q2.2016 (T = 31). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Figure 7: Real GDP growth (one-year rolling horizon forecast)



Note: The four top plots report the realised real GDP growth at the current release (solid black line) along with the median of the forecast distributions (solid pink line), and the 15% and 85% quantiles of the forecast distributions (the pink dotted lines). The bottom plot reports the Ranked Probability Score at the current release for the SPF forecasts (light blue solid line), the Gaussian random walk forecasts (red dashed line) and the naive forecasts (yellow dotted line). Forecasts are for 2001.Q3 to 2017.Q4. The blue vertical line in 2009.Q1 separates the two sub-samples considered.

Table 9: DMW test for the real GDP growth. Full sample Q1.2001 - Q2.2016, $T = 62$.

Ranked Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	2.67	2.57	2.74	0.01	1.61	1.37
WCE-B	2.38**	2.18*	2.46**	0.01	1.80	1.67
WPE-D	2.02*	1.75	2.21*	0.01	1.53	1.74
			Fifth release			
WCE-DM	2.49	2.56	2.65	0.04	1.68	1.37
WCE-B	2.18*	2.18*	2.38**	0.05	1.89	1.68
WPE-D	1.82	1.75	2.13*	0.04	1.61	1.73
			Current release			
WCE-DM	1.94	2.55	2.90	-0.13	1.96	1.02
WCE-B	1.77	2.21*	2.63**	-0.15	2.30**	1.39
WPE-D	1.55	1.77	2.34**	-0.15	2.03*	1.48

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	1.34	2.33	1.79	-0.50	1.22	0.90
WCE-B	1.40	2.27*	1.87	-0.50	1.23	0.98
WPE-D	1.15	1.78	1.74	-0.38	0.96	1.45
			Fifth release			
WCE-DM	1.05	2.07	1.52	-0.47	1.26	1.08
WCE-B	1.04	1.95*	1.53	-0.48	1.29	1.19
WPE-D	0.86	1.53	1.39	-0.37	1.01	1.77
			Current release			
WCE-DM	0.22	1.59	2.14	-0.83	1.10	0.96
WCE-B	0.25	1.62	2.35**	-0.90	1.19	1.13
WPE-D	0.21	1.28	2.30*	-0.72	0.97	1.96*

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the real GDP growth rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the full sample Q1.2001 - Q2.2016 (T = 62). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Table 10: DMW test for the real GDP growth. Sub-sample Q1.2001 - Q3.2008, $T = 31$.

Ranked Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	0.58	1.27	2.08	-1.31	0.52	1.14
WCE-B	0.58	1.26	2.56**	-1.51	0.57	1.20
WPE-D	0.63	1.02	1.93	-1.46	0.51	0.95
			Fifth release			
WCE-DM	0.20	1.29	2.02	-1.25	0.69	1.01
WCE-B	0.20	1.30	2.48**	-1.44	0.77	1.07
WPE-D	0.22	1.05	1.88	-1.36	0.70	0.84
			Current release			
WCE-DM	-0.06	1.28	2.19	-1.23	1.35	0.88
WCE-B	-0.06	1.33	2.66**	-1.47	1.56	0.93
WPE-D	-0.06	1.08	2.06*	-1.41	1.39	0.74

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	-0.30	0.19	0.74	-2.10	-1.13	2.45
WCE-B	-0.33	0.22	0.98	-2.35*	-1.22	1.91
WPE-D	-0.32	0.22	0.88	-2.40*	-1.18	1.81
			Fifth release			
WCE-DM	-0.79	-0.13	0.40	-1.95	-0.83	2.42
WCE-B	-0.89	-0.15	0.55	-2.24*	-0.96	1.98
WPE-D	-0.81	-0.14	0.55	-2.29*	-0.94	1.92
			Current release			
WCE-DM	-0.99	-0.06	0.93	-1.70	-0.34	1.76
WCE-B	-1.09	-0.07	1.34	-1.91	-0.49	1.89
WPE-D	-0.91	-0.06	1.19	-1.83	-0.36	1.86

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the real GDP growth rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the sub-sample Q1.2001 - Q3.2008 (T = 31). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Table 11: DMW test for the real GDP growth. Sub-sample Q4.2008 - Q2.2016, $T = 31$.

Ranked Probability Score

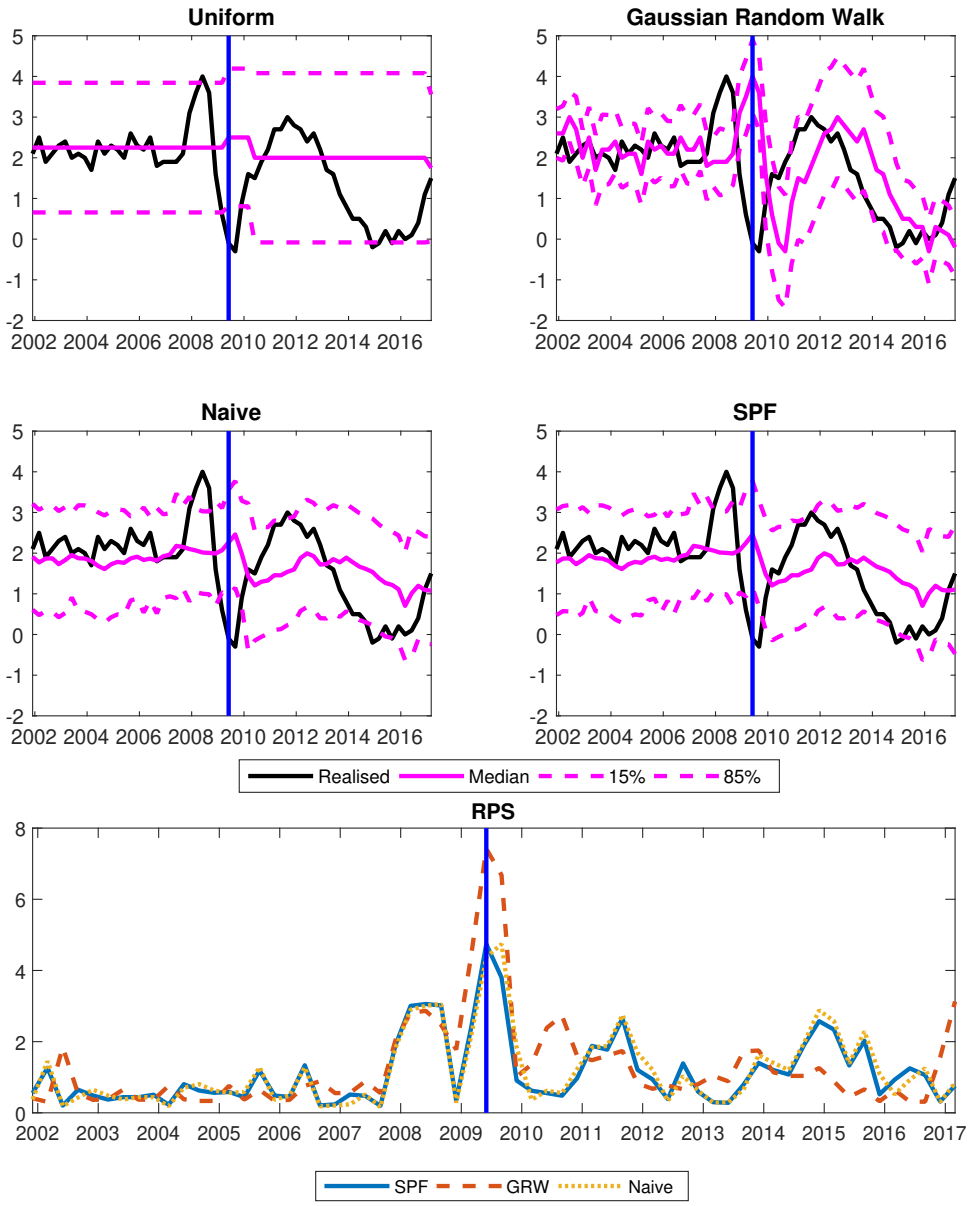
	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	3.20	2.49	2.06	0.94	1.66	0.63
WCE-B	3.13**	2.33*	1.90	1.03	1.88	0.69
WPE-D	2.80**	2.07*	1.70	0.89	1.53	0.57
			Fifth release			
WCE-DM	3.25	2.45	1.99	0.89	1.66	0.74
WCE-B	3.16**	2.30*	1.84	0.97	1.88	0.81
WPE-D	2.84**	2.05*	1.65	0.84	1.54	0.70
			Current release			
WCE-DM	2.47	2.44	2.22	0.75	1.74	0.49
WCE-B	2.47**	2.30*	2.09*	0.84	1.99	0.56
WPE-D	2.21*	2.04*	1.86	0.69	1.60	0.49

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	3.02	3.72	1.71	0.99	2.25	-0.08
WCE-B	2.89**	3.58**	1.67	1.04	2.49**	-0.09
WPE-D	2.61**	3.19**	1.63	0.94	1.96*	-0.08
			Fifth release			
WCE-DM	3.03	3.31	1.61	0.95	2.19	0.17
WCE-B	2.86**	3.24**	1.56	1.00	2.40*	0.20
WPE-D	2.54**	2.87**	1.52	0.90	1.91	0.19
			Current release			
WCE-DM	1.53	2.29	2.06	0.52	1.71	0.03
WCE-B	1.76	2.40*	2.12*	0.56	1.90	0.04
WPE-D	1.59	2.11*	1.95*	0.46	1.43	0.04

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the real GDP growth rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the sub-sample Q4.2008 - Q2.2016 (T = 31). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Figure 8: Inflation (one-year rolling horizon forecast)



Note: The four top plots report the realised inflation at the current release (solid black line) along with the median of the forecast distributions (solid pink line), and the 15% and 85% quantiles of the forecast distributions (the pink dotted lines). The bottom plot reports the Ranked Probability Score at the current release for the SPF forecasts (light blue solid line), the Gaussian random walk forecasts (red dashed line) and the naive forecasts (yellow dotted line). Forecasts are for December 2001 to March 2017. The blue vertical line in June 2009 separates the two sub-samples considered.

Table 12: DMW test for the HICP. Full sample Q1.2001 - Q2.2016, $T = 62$.

Ranked Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	1.05	0.90	1.26	0.07	1.41	1.48
WCE-B	0.92	0.80	1.53	0.07	1.41	1.27
WPE-D	0.87	0.77	1.31	0.07	1.31	1.06
			Fifth release			
WCE-DM	1.17	0.89	1.40	0.08	1.49	1.51
WCE-B	1.07	0.79	1.64	0.09	1.49	1.29
WPE-D	1.04	0.76	1.36	0.08	1.36	1.08
			Current release			
WCE-DM	0.99	0.94	1.32	0.10	1.40	1.45
WCE-B	0.86	0.83	1.55	0.11	1.40	1.24
WPE-D	0.81	0.79	1.31	0.10	1.30	1.03

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	-0.55	-1.04	-0.10	-0.08	0.41	1.48
WCE-B	-0.53	-1.04	-0.11	-0.08	0.46	0.85
WPE-D	-0.57	-1.00	-0.11	-0.07	0.46	0.66
			Fifth release			
WCE-DM	-0.31	-0.87	0.30	-0.11	0.44	1.76
WCE-B	-0.30	-0.86	0.31	-0.11	0.48	1.40
WPE-D	-0.33	-0.82	0.25	-0.09	0.43	1.21
			Current release			
WCE-DM	-0.60	-1.06	0.06	0.02	0.48	0.90
WCE-B	-0.58	-1.06	0.06	0.02	0.54	0.62
WPE-D	-0.60	-1.02	0.05	0.01	0.56	0.48

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the inflation rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the full sample Q1.2001 - Q2.2016 (T = 62). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Table 13: DMW test for the HICP. Sub-sample Q1.2001 - Q3.2008, $T = 31$.

Ranked Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	-0.93	1.07	-1.47	-0.68	0.47	0.15
WCE-B	-0.87	1.11	-1.45	-0.75	0.53	0.14
WPE-D	-0.76	1.10	-1.27	-0.65	0.45	0.13
			Fifth release			
WCE-DM	-0.79	1.06	-1.30	-0.72	0.63	0.22
WCE-B	-0.74	1.10	-1.30	-0.79	0.70	0.20
WPE-D	-0.66	1.09	-1.12	-0.70	0.59	0.19
			Current release			
WCE-DM	-0.93	1.07	-1.47	-0.56	0.45	0.12
WCE-B	-0.87	1.11	-1.45	-0.62	0.51	0.11
WPE-D	-0.76	1.10	-1.27	-0.54	0.43	0.10

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	-0.03	-0.46	-1.30	0.13	0.20	0.20
WCE-B	-0.03	-0.47	-1.33	0.14	0.22	0.19
WPE-D	-0.02	-0.47	-1.20	0.13	0.18	0.19
			Fifth release			
WCE-DM	0.14	-0.29	-1.35	0.08	0.24	0.87
WCE-B	0.12	-0.29	-1.38	0.09	0.26	0.83
WPE-D	0.11	-0.29	-1.24	0.08	0.22	0.75
			Current release			
WCE-DM	-0.03	-0.46	-1.30	0.26	0.32	-0.05
WCE-B	-0.03	-0.47	-1.33	0.28	0.36	-0.04
WPE-D	-0.02	-0.47	-1.20	0.25	0.30	-0.05

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the inflation rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the sub-sample Q1.2001 - Q3.2008 (T = 31). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

Table 14: DMW test for the HICP. Sub-sample Q4.2008 - Q2.2016, $T = 31$.

Ranked Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	2.08	0.45	1.92	0.58	1.49	1.79
WCE-B	1.86	0.42	2.28*	0.63	1.57	1.68
WPE-D	1.66	0.34	2.31*	0.53	1.38	1.57
			Fifth release			
WCE-DM	2.12	0.43	2.02	0.62	1.49	1.79
WCE-B	1.95	0.41	2.27*	0.67	1.57	1.68
WPE-D	1.73	0.33	2.24*	0.56	1.38	1.57
			Current release			
WCE-DM	2.00	0.50	2.00	0.56	1.49	1.79
WCE-B	1.78	0.47	2.31*	0.61	1.57	1.68
WPE-D	1.59	0.38	2.33*	0.51	1.38	1.57*

Quadratic Probability Score

	1 year ahead			2 years ahead		
	Uniform	GRW	Naive	Uniform	GRW	Naive
			Initial release			
WCE-DM	-0.77	-0.94	1.08	-0.27	0.39	1.73
WCE-B	-0.78	-0.95	1.56	-0.29	0.41	1.17
WPE-D	-0.64	-0.79	2.03*	-0.23	0.34	1.18
			Fifth release			
WCE-DM	-0.61	-0.85	1.42	-0.27	0.39	1.73
WCE-B	-0.62	-0.85	1.71	-0.29	0.41	1.17
WPE-D	-0.50	-0.70	1.66	-0.23	0.34	1.18
			Current release			
WCE-DM	-0.86	-0.97	1.47	-0.27	0.39	1.73
WCE-B	-0.87	-0.98	1.78	-0.29	0.41	1.17
WPE-D	-0.70	-0.81	2.04*	-0.23	0.34	1.18

Note: the table reports DMW test statistic values for one-year and two-year ahead ECB SPF density forecasts for the inflation rate against the uniform, the Gaussian random walk and the naive benchmark forecasts for the sub-sample Q4.2008 - Q2.2016 (T = 31). A negative sign implies that benchmarks perform better than the ECB SPF. Long run variances are estimated using WCE with rectangular kernel and Diebold and Mariano (1995) bandwidth (WCE-DM), WCE with the Bartlett kernel and bandwidth $M = \lfloor T^{1/2} \rfloor$ (WCE-B) and WPE with Daniell kernel and bandwidth $m = \lfloor T^{1/3} \rfloor$ (WPE-D). and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics. Rejections using fixed-smoothing asymptotics are reported using ** and * to indicate, respectively, two-sided significance at the 5% and 10% level.

after 2011 inflation is less close to the target, oscillating between 0% and 3%. Thus, there is little scope for the SPF to outperform the simple benchmarks in the first half of the sample, as both the SPF, the benchmarks and the realisation of inflation are close to the 2% target. On the other hand, it is very interesting to see if the professional forecasters beat the simple benchmarks in the second half of the sample: indeed, in view of the outcomes of the DMW tests for unemployment and real GDP growth, for which we found that professional forecasters did outperform the benchmarks, we may expect the same to happen in the case of HICP.

The outcome of the DMW tests for inflation is in tables 12-14: we find that there is no statistical evidence that ECB SPF density forecasts outperform the benchmarks in the full sample, as the null hypothesis of equal forecast accuracy is never rejected. As we anticipated, there is little scope for the professional forecasters to outperform the simple benchmarks in the first part of the sample, because inflation kept close to the target. The evidence is more nuanced for the second subsample, as for this period we do indeed find some evidence in favour of the SPF forecasts, at least against the naive benchmark. The statistical evidence is rather weak, as we only find significance at 10% level (using fixed-smoothing asymptotics), but this result supports the anecdotal evidence of a change in the forecasting practice after the financial crisis: when inflation is more loosely anchored to the target, professional forecasters take this fact into account, in line with de Vincent-Humphreys, Dimitrova, Falck and Henkel (2019) and the empirical investigations of Lyziak and Paloviita (2017) and Grishchenko, Mouabbi and Renne (2019). However, we notice that the significance for HICP is weaker than the evidence for unemployment and real GDP growth, and in figure 8 we can see that SPF inflation forecasts do move with inflation but remain much closer to the target, signalling that even in this phase the inflation targeting still informs the forecasters.

Overall, using the QPS yields qualitatively the same outcome as using the RPS, although significant results for the DMW tests are less frequent when the QPS is used, see

for example the case of the real GDP growth. The only striking difference is for the case of inflation, where using the RPS we find marginally significant evidence suggesting that in the second subsample professional forecasters outperform at least the naive benchmark: this result does not hold when the QPS is used instead. We do not think that this result weakens the evidence that we obtain using the RPS, but rather that it complements it. As the difference is due to the different loss functions, this indicates that part of the superior predictive ability of the ECB SPF is indeed due to the fact that they can at least place more probability in the neighbourhood of the effective outcome, often near-missing the true realization.

The vintage of the realised value has no effect on HICP forecast evaluation, but there is a minor effect on the unemployment rate and real GDP growth forecasts: using the initial release we find a stronger evidence that the professional forecasters outperform the simple benchmarks in some cases. This interesting result would be obscured if we used the most recent vintage. As for the benchmarks, the ECB SPF often outperforms the naive benchmark, indicating that forecasters update the forecast using new information. On the other hand, the uniform benchmark seems the most difficult to beat especially for two-years ahead forecasts, for which the test statistic is often negative and, for real GDP growth in the first subsample, even marginally significant, indicating that in this period the uniform benchmark was significantly outperforming the ECB SPF.

Comparing the application of standard asymptotics with fixed-smoothing asymptotics, we reject the null of equal predictive ability more frequently for the tests with standard asymptotics, especially in the subsamples and for long-horizon forecasts. This is due to the fact that in the subsamples the tests are performed only on 31 observations, exacerbating the size distortions induced by standard asymptotics, see Section 6. For example, Table 8 shows that for unemployment using the RPS loss function both the test with WCE and standard asymptotics reject at 5% significance level the null of equal predictive ability of the SPF and the naive benchmark, on the last subsample for the

two-years ahead horizon. The DMW test using the WCE and standard asymptotics also rejects the null at 10% significance level. This could be interpreted as a clear indication of superior predictive ability of the SPF over the naive forecast for the unemployment rate at two-years horizon. However, these results are partially spurious and demonstrate the risks of using standard asymptotics in a small sample. Indeed when using fixed-smoothing asymptotics, we only find limited evidence of superior predictive ability of the ECB SPF with respect to the naive benchmark at two-years horizon.

It is interesting to compare our results with findings in the extant literature. Our interpretation of our findings about inflation forecasts is consistent with the discussion in Lyziak and Paloviita (2017) and Grishchenko, Mouabbi and Renne (2019), but our empirical investigation rather has aspects in common with Kenny, Kostka and Masera (2014), including the choice of the same forecast rules used as simple benchmarks. The analysis in Kenny, Kostka and Masera (2014) is more disaggregated, but covers a shorter sample (ending with 2011) and uses standard asymptotics. Their main findings for real GDP growth and HICP inflation are broadly consistent with the results that we obtain in the first subsample. Extending the sample and using fixed-smoothing asymptotics, however, gives us the opportunity to study a period with more variability in real GDP growth and unemployment, as well as in inflation: this may cause larger forecast errors and make forecasting more challenging but also more important, as it allows us to detect a change in the forecasting practice after the financial crisis.

8 Conclusions

We analyse ECB SPF aggregate density forecasts over the period 2001.Q1-2016.Q2 using the Diebold-Mariano-West test for equal forecast accuracy with fixed-smoothing asymptotics, taking as benchmarks forecasts generated from a uniform distribution, a Gaussian random walk and the previous survey round.

Our results indicate that ECB SPF density forecasts for unemployment and real

GDP growth outperform the benchmarks, especially at one-year ahead. We also find that survey forecasts for inflation do not easily outperform the benchmarks. For all the variables, however, we find evidence of an improvement in predictive ability since 2009, supporting the anecdotal evidence of a change in the forecasting practice after the financial crisis.

A Example Monte Carlo

In this appendix, we show directly that the process d_t in Example 1 satisfies the FCLT.

Notice that $\sigma_T^2 = 1/2\,(Var(d_1) + Var(d_2) + Cov(d_1, d_2))$. When $\lfloor rT \rfloor$ is even we rearrange

$$\sum_{t=1}^{\lfloor rT \rfloor} d_t = (d_1 + d_2) + (d_3 + d_4) + \cdots + (d_{\lfloor rT \rfloor - 1} + d_{\lfloor rT \rfloor}) = D_1 + D_2 + \cdots + D_{\lfloor rS \rfloor}$$

where $S = T/2$ and notice that $D_s = (d_{2s-1} + d_{2s})$ is a sequence of independently, identically distributed variables with variance $\sigma_S^2 = Var(D_s) = 2\sigma_T^2$. When $\lfloor rT \rfloor$ is odd we rearrange

$$\sum_{t=1}^{\lfloor rT \rfloor} d_t = D_1 + D_2 + \cdots + D_{\lfloor rS \rfloor} + d_{\lfloor rT \rfloor}$$

and notice that $d_{\lfloor rT \rfloor}$ is $O_p(1)$ uniformly in r because d_t has bounded support by construction. Thus, uniformly in r, it holds that

$$\sqrt{T}/T \; 1/\sigma_T \sum_{t=1}^{\lfloor rT \rfloor} d_t = \sqrt{S}/S \; 1/\sigma_S \sum_{s=1}^{\lfloor rS \rfloor} D_s + O_p(1/\sqrt{T}) \Rightarrow W(r)$$

as $T \to \infty$.

References

Andrews, Donald W. K. (1984) 'Non-strong mixing autoregressive processes.' *Journal of Applied Probability* 21(4), 930–934

Boero, Gianna, Jeremy Smith, and Kenneth F Wallis (2008) 'Uncertainty and disagreement in economic prediction: the Bank of England Survey of External Forecasters.' *The Economic Journal* 118(530), 1107–1127

Bowles, Carlos, Roberta Friz, Veronique Genre, Geoff Kenny, Aidan Meyler, and Tuomas Rautanen (2007) 'The ECB Survey of Professional Forecasters (SPF)-a review after eight years' experience.' *ECB Occasional Paper*

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

Clark, Todd E (1999) 'Finite-sample properties of tests for equal forecast accuracy.' *Journal of Forecasting* 18(7), 489–504

Clements, Michael P (2014) 'Forecast uncertainty—ex ante and ex post: US inflation and output growth.' *Journal of Business and Economic Statistics* 32(2), 206–216

Coroneo, Laura, and Fabrizio Iacone (2015) 'Comparing predictive accuracy in small samples.' Technical Report 15/15, The University of York

D'Amico, Stefania, Athanasios Orphanides et al. (2008) 'Uncertainty and disagreement in economic forecasting.' *Federal Reserve Board Finance and Economics Discussion Series*

Dawid, A Philip (1984) 'Present position and potential developments: Some personal views statistical theory the prequential approach.' *Journal of the Royal Statistical Society: Series A (General)* 147(2), 278–290

de Vincent-Humphreys, Rupert, Ivelina Dimitrova, Elisabeth Falck, and Lukas Henkel (2019) 'Twenty years of the ECB survey of professional forecasters.' *Economic Bulletin Articles*

Diebold, Francis X, and Robert S Mariano (1995) 'Comparing predictive accuracy.' *Journal of Business and Economic Statistics* 13(3), 253–262

Diebold, Francis X., Anthony S. Tay, and Kenneth F. Wallis (1997) 'Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters.' NBER Working Papers 6228, National Bureau of Economic Research, Inc, October

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

Fair, Ray C (1980) 'Estimating the expected predictive accuracy of econometric models.' *International Economic Review* 21(2), 355–378

Garcia, Juan A (2003) 'An introduction to the ECB's Survey of Professional Forecasters.' *ECB Occasional Paper*

Geraats, Petra M (2008) 'ECB credibility and transparency.' *Economic Papers* 330, 1–33

Giannone, Domenico, Jérôme Henry, Magdalena Lalik, and Michele Modugno (2012) 'An area-wide real-time database for the Euro area.' *Review of Economics and Statistics* 94(4), 1000–1013

Gneiting, Tilmann, and Adrian E Raftery (2007) 'Strictly proper scoring rules, prediction, and estimation.' *Journal of the American Statistical Association* 102(477), 359–378

Grishchenko, Olesya, Sarah Mouabbi, and Jean-Paul Renne (2019) 'Measuring inflation anchoring and uncertainty: A U.S. and Euro Area comparison.' *Journal of Money, Credit and Banking* 51(5), 1053–1096

Hartmann, Philipp, and Frank Smets (2018) 'The first twenty years of the European Central Bank: monetary policy.' *ECB Working Paper*

Harvey, David I., Stephen J. Leybourne, and Emily J. Whitehouse (2017) 'Forecast evaluation tests and negative long-run variance estimates in small samples.' *International Journal of Forecasting* 33(4), 833 – 847

Henckel, Timo, Gordon D Menzies, Peter Moffatt, and Daniel J Zizzo (2019) 'Three dimensions of central bank credibility and inferential expectations: The euro zone.' *Journal of Macroeconomics* 60, 294–308

Hualde, Javier, and Fabrizio Iacone (2017) 'Fixed bandwidth asymptotics for the studentized mean of fractionally integrated processes.' *Economics Letters* 150, 39–43

Kenny, Geoff, Thomas Kostka, and Federico Masera (2014) 'How informative are the subjective density forecasts of macroeconomists?' *Journal of Forecasting* 33(3), 163–185

Kiefer, Nicholas M, and Timothy J Vogelsang (2002a) 'Heteroskedasticity–autocorrelation robust standard errors using the bartlett kernel without truncation.' *Econometrica* 70(5), 2093–2095

_ (2002b) 'Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size.' *Econometric Theory* 18(6), 1350–1366

_ (2005) 'A new asymptotic theory for heteroskedasticity-autocorrelation robust tests.' *Econometric Theory* 21(6), 1130–1164

Lazarus, Eben, Daniel J. Lewis, James H. Stock, and Mark W. Watson (2018) 'HAR inference: Recommendations for practice.' *Journal of Business and Economic Statistics* 36(4), 541–559

Li, Jia, and Andrew J. Patton (2018) 'Asymptotic inference about predictive accuracy using high frequency data.' *Journal of Econometrics* 203(2), 223 – 240

Lyziak, Tomasz, and Maritta Paloviita (2017) 'Anchoring of inflation expectations in the euro area: Recent evidence based on survey data.' *European Journal of Political Economy* 46, 52 – 73

Mankiw, N Gregory, and Matthew D Shapiro (1986) 'News or noise? An analysis of GNP revisions.' Technical Report, National Bureau of Economic Research Cambridge, Massachusetts, USA

Manzan, Sebastiano (2017) 'Are professional forecasters bayesian?' Technical Report, Zicklin School of Business, Baruch College, CUNY

Mitchell, James, and Stephen G Hall (2005) 'Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and niesr fan charts of inflation.' *Oxford Bulletin of Economics and Statistics* 67, 995–1033

Neave, Henry R (1970) 'An improved formula for the asymptotic variance of spectrum estimates.' *The Annals of Mathematical Statistics* 41(1), 70–77

Newey, Whitney K., and Kenneth D. West (1994) 'Automatic lag selection in covariance matrix estimation.' *The Review of Economic Studies* 61(4), 631–653

Phillips, Peter C. B. (2005) 'HAC estimation by automated regression.' *Econometric Theory* 21(1), 116–142

Phillips, Peter C. B., and Victor Solo (1992) 'Asymptotics for linear processes.' *Annals of Statististics* 20(2), 971–1001

Stark, Tom (2010) 'Realistic evaluation of real-time forecasts in the survey of professional forecasters.' *Federal Reserve Bank of Philadelphia Research Rap, Special Report*

Sun, Yixiao (2013) 'A heteroskedasticity and autocorrelation robust F test using an orthonormal series variance estimator.' *The Econometrics Journal* 16(1), 1–26

Sun, Yixiao (2014) 'Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference.' *Journal of Econometrics* 178, 659 – 677

Tay, Anthony S, and Kenneth F Wallis (2000) 'Density forecasting: a survey.' *Journal of Forecasting* 19(4), 235–254

West, Kenneth D (1996) 'Asymptotic inference about predictive ability.' *Econometrica* 64(5), 1067–1084

Wooldridge, Jeffrey M., and Halbert White (1988) 'Some invariance principles and central limit theorems for dependent heterogeneous processes.' *Econometric Theory* 4(2), 210–230