EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Georgios Kontogeorgos, Kyriacos Lambrias

An analysis of the Eurosystem/ECB projections

No 2291 / June 2019

Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Abstract

The Eurosystem/ECB staff macroeconomic projection exercises constitute an important input to the ECB's monetary policy. This work marks a thorough analysis of the Eurosystem/ECB projection errors by looking at criteria of optimality and rationality using techniques widely employed in the applied literature of forecast evaluation. In general, the results are encouraging and suggest that Eurosystem/ECB staff projections abide to the main characteristics that constitute them reliable as a policy input. Projections of GDP - up to one year - and inflation are optimal - in the case of inflation they are also rational. A main finding is that GDP forecasts can be substantially improved, especially at long horizons.

Keywords: Eurosystem/ECB forecasts, Forecast evaluation, Forecast errors

JEL Classification: C53, E37, E58

Non-technical summary

The Eurosystem/ECB staff macroeconomic projections, officially known as (Broad) Macroeconomic Projection Exercises or (B)MPEs, are an essential input to the conduct of the common monetary policy in the euro area. Thereby, it is important that these projections abide to certain characteristics that would constitute them credible both to the policy maker and to the public. This paper makes a step towards this direction in evaluating the forecasting performance of the (B)MPEs with respect to the two key variables: real GDP growth and HICP inflation for the euro area as a whole. The analysis focuses on the projections of these variables mainly at one-quarter, one-year and two-year projection horizon over the period 2001Q4 - 2016Q3.

We perform statistical tests to evaluate the optimality, efficiency and, consequently, rationality of the Eurosystem/ECB projections. Forecasts are optimal when they minimise a given loss function. They are (strongly) efficient when they fully take into account all publicly available information at the time a forecast was made. Finally, they are rational when they satisfy both optimality and efficiency. These properties translate into testable statistical analysis that we implement on the Eurosystem/ECB projections of GDP growth and HICP inflation in year-on-year terms (y-o-y) at quarterly frequency. Further, we go beyond the theoretical properties and provide a series of standard statistical tests to evaluate the reliability and accuracy of the Eurosystem/ECB projections.

Overall, our analysis gives credit to the (B)MPE forecaster, while pointing to some areas of potential improvement especially with regards to GDP projections. We find no evidence of systematic bias in the projections of GDP growth and inflation, with only exception the projections of GDP growth at horizons close to two years (tendency to over-predict). Further, we find that the forecasts are weakly efficient: we document the existence of an autocorrelation in the forecast error that is aligned with the theory - except in one case (GDP at one-quarter horizon). For the case of inflation, we find that forecasts are “strongly efficient”, but the same conclusion cannot be derived for the GDP projections. Finally, we also show that the forecasting accuracy of both variables deteriorates with the forecast horizon. Overall, GDP and inflation forecasts are *optimal* - except for GDP at long-horizons - and in the case of inflation *rational*.

We also show that inflation forecast errors are normally distributed but GDP errors are not. The path of the variables has been in general correctly anticipated - with some exceptions - and the Eurosystem/ECB forecasts feature relatively well against simple forecasting benchmark models and against other institutions and private-sector forecasts.

Overall, a main outcome of the analysis is to suggest that the long-term GDP forecasts can be substantially improved.

1 Introduction

The Eurosystem/ECB staff macroeconomic projections, officially known as (Broad) Macroeconomic Projection Exercises or (B)MPEs, are an essential input to the conduct of the common monetary policy in the euro area. Thereby, it is important that these projections abide to certain characteristics that would constitute them credible both to the policy maker and to the public. This paper makes a step towards this direction in evaluating the forecasting performance of the (B)MPEs (henceforth Eurosystem/ECB projections and (B)MPEs are used interchangeably) with respect to the two key variables: real GDP growth and HICP inflation for the euro area as a whole. The analysis focuses on the projections of these variables mainly at one-quarter, one-year and two-year projection horizon over the period 2001Q4 - 2016Q3.

This paper adds to the existing analyses of the Eurosystem/ECB projections - like ECB (2012), ECB (2013) and Alessi et al. (2014) - and it employs tests and forecast evaluation methods common in the literature to derive robust conclusions about the quality of the (B)MPEs. In this regard, it shares a lot of elements to analyses and evaluation checks of the forecasts of other organisations conducted in an institutional context, such as commissioned studies or by the respective institutions' own staff. A non-exhaustive list of these analyses include Melander et al. (2007) and Fioramanti et al. (2016) for the European Commission (EC), IMF-IEO (2014) for the IMF[1], Vogel (2007) and Pain et al. (2014) for the OECD, and the BoE-IEO (2015) for the Bank of England. It also shares, and uses, elements from the academic literature of forecast evaluation of policy institutions and other professional forecasts, such as Romer and Romer (2000), Ager et al. (2009), Clements et al. (2007) and El-Shagi et al. (2016).

We perform statistical tests to evaluate the optimality, efficiency and, consequently, rationality of the Eurosystem/ECB forecasts. Forecasts are optimal when they minimise a given loss function. The current analysis is conducted under the standard setting in the forecasting literature of a *quadratic* loss function, i.e. Mean Squared Error (MSE) loss, $L(e) = \alpha e^2$, where e is the forecast error[2]. Under the assumed squared loss function, *optimal* forecasts satisfy the

[1] See Freedman (2014) for an overview of all commissioned studies at the IMF.

[2] This function is symmetric, differentiable everywhere and penalises large forecast errors at an increasing rate due to its convexity in $|e|$ - Elliot and Timmermann (2016, p.20). Draghi (2016a, 2016b) has referred to the symmetric definition of the objective of price stability, however naturally it is open to question whether the Eurosystem/ECB has, or should have, a symmetric and quadratic loss function for all the forecasted variables. For a more generic treatment of the optimal forecast error properties under unknown loss see Elliot et al. (2005) and Patton and Timmermann (2007a, 2007b).

following properties (Elliot and Timmermann, 2016, Section 15.3.1):

1. Unbiasedness - expectation of forecast error for period $t+h$ done at period t, where h is the forecast horizon, should be zero at t, both conditionally on the information set that was available to the forecaster (Ω_t) and unconditionally:

$$E[e_{t+h}|\Omega_t] = E[e_{t+h}] = 0 \tag{1}$$

2. Weak efficiency - the forecast error behaves as a $MA(h-1)$ process:

$$cov(e_{t+h}, e_{t-j}) = 0, \;\; \forall j \geq 0 \tag{2}$$

3. Variance of the forecast error is a non-decreasing function of the forecast horizon h:

$$var(e_{t+h}) \leq var(e_{t+h+1}), \;\; \forall h \tag{3}$$

Furthermore, forecasts are *rational* when they are optimal and (strongly) efficient. Efficient forecasts fully account for all available information that was at the disposal of the forecaster at the time the forecast was made. Therefore, strong efficiency refers to extending the forecasters information set to include not only past outcomes and forecasts (and hence past forecast errors) of the forecasted variable under question, but to include all other variables that were publicly available at the time of the forecast (Elliot and Timmermann, 2016, p. 356).

These properties translate into testable statistical analysis that we implement on the Eurosystem/ECB forecasts of GDP growth and HICP inflation in year-on-year terms (y-o-y) at quarterly frequency. Further, we go beyond the theoretical properties and provide a series of standard statistical tests to evaluate the reliability and accuracy of the Eurosystem/ECB forecasts. Normality of the forecast errors, although not an optimal property, it is quite often implicitly or explicitly assumed so that multiples of standard deviation are used to derive confidence bands around the point forecasts. Standardised third and fourth moments, i.e. the skewness and kurtosis are tested separately and then jointly to conclude on whether there is evidence for the normality of the (B)MPE errors. We also test the directional accuracy of the Eurosystem/ECB staff forecasts. That is, we test on whether the direction of the projected variables follows the realised path. This is of course very important for policy analysis. Finally, we test the (B)MPEs forecasting performance against simple benchmark-models - like the Random Walk (RW) and the autoregressive process of order one - AR(1) - and against other forecasters. For many of the tests we do, we check how the performance has changed through time and how much it has been influenced by errors in the conditioning assumptions and the financial crisis.

Overall, the main results of the analysis give credit to the (B)MPE forecaster, while pointing to some areas of potential improvement, especially with regards to GDP forecasts.

Unbiasedness: We find no evidence of systematic bias in the forecasts of GDP growth and inflation, with only exception the projection of GDP growth at horizons close to two years (tendency to over-predict). Unbiasedness is heavily scrutinised in our analysis, as it is an important property for the credibility of any forecast and it is in general easily picked-up by the "watchers" of the Eurosystem/ECB, like financial analysts and the financial press. We start by evaluating the existence of a bias over the whole sample and at each forecast horizon separately, as it is usual in the literature, using a combination of Mincer and Zarnowitz (MZ, 1969) regressions together with the Holden and Peel (HP, 1990) tests. Based on the combined results from both tests we can confidently conclude the lack of bias in the inflation forecasts and the bias in the GDP forecasts at long horizons. In a second step, we evaluate the bias by pooling information across several forecasting horizons, following the approach suggested by Clements et al. (2007) and Ager et al. (2009). We find no evidence of a common bias across all horizons for both GDP and inflation forecasts. We do find, however, strong evidence of horizon-specific bias once we allow for it. This is captured by calculating the bias at each forecast horizon, similar to the common approach with HP regressions, but testing whether these biases are *jointly* equal to zero. In that case, we find that GDP forecasts are biased already at 2-quarters ahead forecasting horizon, and we also find bias in the inflation forecasts at long horizons. Putting the pieces together, we conclude that the inflation forecasts are unbiased as they fail only one out of the three tests we perform, and indeed only at long horizons. Further, we conclude that GDP forecasts up to one-year ahead are not biased, as only one of the three tests suggests otherwise, but at the same time we find strong conclusive evidence of a bias in the long-term GDP forecasts (close to two-years).

Efficiency: Further, we find that the forecasts are weakly efficient: we document the existence of an autocorrelation in the forecast error of order up to the forecast horizon minus one - except in one case (GDP at one-quarter horizon). For the case of inflation, we find that forecasts are "strongly efficient", as they account for available information at the time of the forecast. In particular, we tested whether inflation forecast errors could be related to inflation and GDP data and forecast errors that were available to the forecaster at the time the forecast was made (Gavin and Mandal 2003) - and found no significant relation. This was not the case for the GDP forecast errors where past data on GDP and inflation, as well as forecast errors of those variables, could explain the GDP forecast error in several occasions and at least once in each forecasting horizon tested - one quarter, one year and two years.

Non-decreasing variance: We show that the forecasting accuracy of both variables deteriorates with the forecast horizon. The standard deviation and the Root Mean Squared Error (RMSE) of the GDP and inflation forecasts are not decreasing with the projection horizon. Furthermore, the 95% confidence intervals of the RMSEs at long horizons - derived with bootstrapping techniques - do not overlap with those at short-horizons. With this, we can be confident that this difference is statistically significant.

With these findings we can conclude that the Eurosystem/ECB macroeconomic projections generally satisfy the theoretical properties of optimal forecasts - regarding the inflation forecasts and GDP forecasts at horizons up to one year. However, optimality clearly fails for long-term GDP forecasts. Further, inflation forecasts are rational as they properly take into account information that were available to the forecaster at the time the forecast was made. This cannot be said for the GDP forecasts.

Concerning other tests of forecasting performance and accuracy, the main results are as follows:

Normality: We find evidence that HICP inflation projection errors are normally distributed while GDP growth projection errors are not. In particular, the departure from normality in the GDP growth errors is due to evidence of leptokurtic distributional properties (fat tails), while they do not exhibit any statistically significant skewness. We explore whether large errors committed during the financial crisis are a reason behind the failure of normality and the fat tails. Tests' results improve somewhat, but departures from normality remain especially at long horizons.

Directional accuracy: GDP and inflation forecasts are in general directionally accurate, as they have correctly anticipated the realised path of the variables most often than not, except for GDP at long horizons. Despite this, we cannot confidently conclude that the direction of change of inflation in the forecasts is significantly correlated to the one actually observed; while we do for GDP forecasts up to one year. At the same time, forecasts of GDP growth at long horizons have anticipated a different path than the one realised (i.e. GDP growth accelerating instead of decelerating) more often than not.

Relative performance: (B)MPE forecasts of GDP growth and inflation in general outperform simple benchmarks like the RW and AR(1). Furthermore, the (B)MPE forecasts of euro area GDP growth and inflation feature well against the same forecasts of other international forecasters.

Overall, the results suggest that the GDP growth projections at long horizons can be substantially improved. Putting all the elements together we notice that GDP growth forecasts at or close to two-years ahead are biased, not informationally efficient (an issue at shorter horizons as well) and the direction of change is most often not correctly anticipated.

Time-varying bias: Moreover, we perform tests in a time-varying context. Regarding the bias, we find that GDP forecasts have been biased across the whole life of the Eurosystem/ECB staff forecasts, but statistically significant biases are reported mainly in the early and latest parts of the sample. The bias in the inflation forecast, on the other hand, shows a clear downward trend: the bias was positive (under-prediction) in the pre-crisis period and has been steadily falling to negative territory (over-prediction) in more recent years. Nevertheless, the bias is found to be statistically significant primarily in the early parts of the sample.

Time-varying relative performance: We also test the time-varying performance of the (B)MPE forecasts against the benchmark models and the Survey of Professional Forecasts (SPF). Although the loss differential of the (B)MPE against the benchmark models was generally negative, suggesting overall higher forecasting accuracy of the (B)MPE, it cannot be concluded that this superior performance was statistically significant - except in the case of inflation forecasts at short horizons. The loss differential of the (B)MPE inflation forecasts against the AR(1) model has been following a downward trend: while it was outperformed by the model before the financial crisis, it has been doing better since. The SPF forecasters have in general outperformed the (B)MPE on GDP forecasts across the sample, and for long-horizon forecasts significantly so especially in recent years. On the other hand, the (B)MPE one-year ahead inflation forecasts have been steadily improving against the SPF; and have been outperforming those since the financial crisis.

Overall, putting all the elements of the time-varying tests together, we observe a continuous improvement of the inflation forecasts: a continuous fall in the bias and an improvement in the forecasting accuracy against the AR(1) and the SPF forecasts. Nevertheless, the recent negative bias - the well-known tendency of the (B)MPE to over-predict inflation in recent years (see

Darvas, 2018) - although not statistically significant it could have had a negative impact on the credibility of the (B)MPE forecasts.

As is the case for other institutions, the Eurosystem/ECB staff projections are conditional forecasts[3]. The staff is asked to provide their best assessment of the evolution of economic activity conditional on the given path of certain important macroeconomic and financial aggregates[4]. In principle, in evaluating the quality of conditional forecasts, one is interested in distinguishing between the properties determined by the quality of the model and the properties determined by the quality of the conditioning assumptions. For these reasons, and to be more "fair" to the (B)MPE forecaster, the analysis conducted below is done also against the error after correcting for the errors in assumptions - called "adjusted error". As it would be anticipated, forecast accuracy improves when the errors in the conditioning assumptions are taken into account: the RMSEs are lower and so is the RMSE ratio of the (B)MPE against benchmark models. The main results outlined above hold, and in most cases improve, when considering the adjusted error - for example there is a notable improvement in directional accuracy of the forecasts. One exception is in the lack of evidence of weak efficiency in GDP forecasts which still holds, even after the adjustment for the errors in the assumptions has been made. Nevertheless, we document some worsening in the unbiasedness results that is difficult to explain. We provide some reasons behind this result, but deeper understanding of it is left for future research.

The next section provides some information on the process and the institutional framework of the Eurosystem/ECB forecasts. Section 3 describes the data, provides the definitions and conventions used and presents some graphical and descriptive analysis of the (B)MPE calendar-year forecast errors. Sections 4 and 5 provide the main analysis of this paper: the former analyses the theoretical properties of optimal and rational forecasts and the latter performs some additional tests of forecast performance and accuracy. Section 7 concludes.

2 Description of the process and the institutional framework of the Eurosystem/ECB forecasts

This section provides a brief description of the Eurosystem/ECB staff forecasts, the institutional framework and the underlying process. This is not an exhaustive description. For further details,

[3] This is the reason why the term projections is used, however, in the text we use the terms "forecasts" and "projections" interchangeably denoting the conditional forecasts.

[4] For more information on the technical assumptions see ECB (2006).

the interested reader is referred to ECB (2016) and Alessi et al. (2014, Section 2.2).

The Eurosystem/ECB staff produces macroeconomic projections for the euro area and the individual countries four times a year. The projections produced in June and December are prepared by Eurosystem experts from both the euro area national central banks (NCBs) and the ECB - these two exercises are referred to as the Broad Macroeconomic Projections Exercises (BMPEs). The projections conducted in March and September are mainly the outcome of the ECB staff projection exercise, which primarily involves ECB staff experts, with NCBs' experts providing the short-term inflation projections. These are referred to as Macroeconomic Projection Exercises (MPEs). In the present study we do not make any distinction between the two exercises and we treat them as one single forecast conducted - by assumption - by the same forecaster. When we refer to the forecasts, we use interchangeably the terms Eurosystem/ECB staff forecasts and (B)MPE. The outcome of the macroeconomic projection exercises conducted by Eurosystem/ECB staff is presented to the Governing Council (a report is also produced) as an input to its monetary policy deliberations. Forecasts for several variables are conducted and can be found on the ECB's website. The forecasts are at quarterly frequency. In both exercises, Eurosystem/ECB staff experts produce forecasts for the individual euro area countries and the euro area, the latter being consistent with the country aggregation.

During a BMPE, the preparation of the macroeconomic projections for the euro area and for the individual euro area countries is undertaken by the Working Group on Forecasting (WGF) under the responsibility and guidance of the Monetary Policy Committee (MPC), which has ownership of the BMPE projection figures and the projection report. During an MPE, the Forecast Task Force (FTF), a group comprising experts from a wide range of business areas within the ECB, is responsible for the production of the projection figures. Guiding the work of the FTF is the Forecast Steering Committee (FSC), which consists of ECB managers. ECB staff is responsible for compiling the resultant MPE report, whose structure is the same as the BMPE report. The MPE report is presented to the MPC, whose Chair conveys the Committee's opinion on the outcome of the exercise in the form of a letter to the President of the ECB.

Eurosystem/ECB staff forecasts are conditional forecasts - conditional on a set of assumptions about the international environment, financial conditions and fiscal variables. Some of the assumptions are derived in a purely technical manner - primarily those concerning financial variables and oil prices. For example, interest rates are assumed to follow market expectations, the exchange rate is assumed to be constant over the forecast horizon, oil prices are derived based on futures price of Brent crude oil etc. Other assumptions are derived by ECB staff and are actually a forecast themselves. These pertain to the international environment, effectively a forecast of the global economy by ECB staff. In this exercise we also check for forecast errors after having accounted for the errors in the assumptions. To do this, we employ the Basic Model Elasticities (BMEs) - a tool developed exactly for checking the impact of changes in the assumptions to a given projection; in a mechanical manner. In essence, BMEs can be thought of as a smaller version of a multi-country model linearised around a specific baseline. The NCBs provide the underlying impulse responses to changes in the exogenous variable(s) for their own countries - e.g. oil prices being 10% higher than in the baseline. ECB staff collects the elasticities from the NCBs and compiles the resulting euro area BMEs. The responsibility and ownership of the tool lays with the WGF. BMEs are updated once per year.

Finally, it is worth emphasising that in both the BMPE and the MPE, NCBs provide short-term forecasts for overall HICP inflation and its key components (unprocessed food, processed food, non-energy industrial goods, energy and services) for their respective countries, with a monthly frequency, over a horizon of 11 months. ECB staff aggregates these individual country inflation figures in order to obtain the euro area inflation path. This is referred to as the Narrow Inflation Projection Exercise (NIPE).

3 Definitions, data and descriptive analysis

3.1 Projection error and adjusted error

We perform tests on the statistical properties of the forecast errors of euro area GDP growth and HICP inflation. For the main part of the analysis, data are year-on-year growth rates at quarterly frequency. When comparing against other institutions we use calendar-year forecasts (annual frequency). The projection error is defined as the realisation minus the projected value. Using the following notation, the h-quarter ahead projection error at quarterly frequency is defined as:

$$e_{th} = y_t^{t+5} - f_{th} \tag{4}$$

where t is the time period being forecasted, e_{th} is the h-horizon forecast error for a variable of interest, y_t^{t+5} stands for the realised value of that variable at time t using the estimate released five quarters ahead of the date of interest $t+5$ and f_{th} is the forecast of that variable for period t produced at period $t-(h-1)$. Calendar-year forecast errors are defined in a similar fashion. The choice of using the "$t+5$" rule is to strike a fair balance between more recent, accurate data and the first release of data which is, on one hand, more in line with the data that were available to the forecaster in real time but, at the same time, subject to significant revisions in forthcoming releases. This choice mostly affects the GDP forecast error but it is rather inconsequential for HICP inflation. Indicatively, other studies in forecast evaluation employing real time datasets follow similar practices using outcomes varying from first releases (El-Shagi et al., 2016), releases two quarters ahead of the reference period (Tulip, 2006; Faust and Wright, 2009; Champagne et al., 2018) up to two years ahead (Faust and Wright, 2008).

The analysis is also conducted for forecast errors adjusted for errors in the conditional assumptions. The current analysis has taken into consideration the following five assumptions: oil prices, foreign demand, exchange rates, long term interest rates and short term interest rates[5]. Once the errors in the assumptions are defined, the adjusted forecasts and consequently the adjusted forecast errors are computed using internal ECB models. In particular, the Basic Model Elasticities (BMEs) are used - see Section 2 and ECB (2016). Regarding short term interest rates, which are of pivotal role from a central bank perspective, some particularly relevant changes and policy measures have taken place over the period under scrutiny. Since June 2006 forecasts are conditional on markets' expectations, whereas before that they were conditional on a flat, unchanged path. Additionally, in July 2013 the ECB introduced forward policy guidance (ECB 2014).

The h-horizon part of the forecast error of the variable of interest explained by errors in assumption i ($\tilde{e}_{thi}$) is given by:

$$\tilde{e}_{thi} = f_{BME}\left(\overrightarrow{e_i}\right) \tag{5}$$

where function f_{BME} represents the function that transforms the vector of the assumption i errors $\overrightarrow{e_i} = [e_{t-(h-1),1,i} ... e_{t,h,i}]$ to the part of the h-horizon forecast error explained by the given assumption i for period t using the BMEs as outlined in the previous section. Note that for a given forecast error of an assumption i at time t, forecast horizon j, $1 \leq \ j \ \leq h$, one needs to account for all the forecast errors up to horizon j. This is because of lagged effects: an erroneous

[5] Data on these assumptions exist for the entire sample analysed in this paper. Further, with the exception of foreign demand, the other assumptions are market-based and involve no model-based analysis in their estimation.

assumption on oil prices is effectively a contemporaneous shock in that period that has impacts potentially for several periods. Furthermore, in order to be as much precise as possible f_{BME} corresponds to the actual BMEs available at the time the forecast was made. The BMEs are re-estimated every year and thus f_{BME} is actually the time-varying function $f_{BME_t}^{EA_i}$ where EA_i refers to the euro area composition of each (B)MPE. The *adjusted error* ($e_{th}^{adjusted}$), which is the item of interest, takes into account errors in all five assumptions. It is thus defined as the difference between the *total* forecast error - as in equation (4) - and the sum of all errors explained by errors in each assumption $i = 1, ..., 5$:

$$e_{th}^{adjusted} = e_{th} - \sum_{i=1}^{5} \tilde{e}_{thi} \tag{6}$$

In the following sections for horizon specific analysis for simplicity of notation we drop subscript that denote the h-horizon series forecast errors, as well as the superscript $t + 5$ vintage data series used as the outcome.

3.2 Data

For each (B)MPE conducted in period t, t refers to the 1^{st} quarter of the projections horizon, whereas the last data point is assumed to be at period $t - 1$. (B)MPEs involve maximum 12 quarters forecasts and in this analysis we perform tests mainly against projection horizons of 1-quarter ($h = 1$, current-quarter forecast), 4-quarters ($h = 4$) and 8-quarters ($h = 8$) forecast errors. More details are provided in Table 16 in the Appendix.

The analysis on full sample concerns the (B)MPEs performed over 2001Q4 - 2016Q3. We choose to focus the analysis on the post-2000 period when the (B)MPE forecasts for the euro area GDP growth and inflation became publicly available. Yet, due to data availability issues of the assumptions, the sample starts in 2001Q4 - for comparability of the analytical results we choose to maintain the same sample size between the two types of forecast error (total and adjusted). We stop our analysis at 2016Q3 to have a consistent definition of the forecast error; in particular of using the vintage time-series released five quarters ahead of the projected value as explained above. Overall, the number of observations varies with the forecast horizon as follows: 60, 57 and 53 observations for $h = 1$, $h = 4$ and $h = 8$ respectively.

When examining the forecast performance over time for the total forecast errors the sample size is expanded backwards from 1999Q1 - 2016Q3 in order to have the maximum possible observations.

3.3 Descriptive analysis

We start the analysis by providing a description of the data, the underlying forecast errors. Figure 1 shows the next calendar-year-horizon forecast errors (yellow bars) produced in the June BMPEs, together with the difference between the actual GDP value for that year and its long-run mean using latest data (blue bars). The actual values themselves are the black horizontal lines for each calendar year and the dot-dash line shows the long run mean. The error bands represent the minimum and maximum forecast errors based on the published forecast ranges, while the forecast error bars themselves correspond to the final outcome minus the mid-point of the forecast range. Consequently, when the range of the forecast errors includes zero it means that the published forecast range included the outcome. The difference between actual GDP and its long-run mean is a simple measure of cyclical economic expansions and downturns from a historical, ex-post perspective. It can be seen that projections have tended to be more optimistic in downturns - i.e. make negative projection errors when the difference between the actual GDP growth rate and its long-run mean was negative - and more pessimistic in expansions. In other words, there has been a tendency to under-predict GDP growth during expansions and over-predict during downturns (in all except of two cases, 2005 and 2014). This pattern for GDP can be seen more clearly in the more detailed Figure 9 in the Appendix. Beyond the pattern, the figure shows that the forecast gets closer to the actual value as the horizon shortens and more information is gathered.

Further, one can see that the financial crisis weighs heavily on the GDP forecast error while the accuracy of GDP forecasts has improved significantly in recent periods (since 2014). Indeed, the RMSE is lower when the errors made during the crisis are not taken into account (see Figure 13 in the Appendix. For example, at one-year horizon, the RMSE is 1.6 overall and 1.0 excluding the financial crisis. The post-crisis period also includes the large forecast errors made during the euro area sovereign-debt crisis, a reason behind the relative poor forecasting performance against the pre-crisis period[6].

Inflation projection errors also show a similar cyclical pattern - there appears to be a tendency to under-predict (over-predict) inflation when actual inflation is above (below) its long-run mean

[6] This does not account for any differences in the sample size over the two periods.

value. This pattern is evident during the pre-crisis years (2001-2008) where there appeared to be a strong tendency to under-predict inflation - while inflation was persistently above its long-run mean - and one after the sovereign debt crisis (2013-2016) where inflation was persistently over-predicted. Paloviita et al. (2017) find that HICP inflation projections exhibit stronger and faster mean reversion compared to the realised inflation. They also find that for inflation forecasts the median projections after about six quarters are already very close to the levels at the end of the forecast horizons. These elements potentially explain the strong cyclical behaviour of the errors. The analysis on unbiasedness of the Eurosystem/ECB inflation forecasts that follows will investigate these patterns more thoroughly. Overall, inflation forecasts appear to be more accurate than the GDP forecasts - especially at short horizons as shown in the RMSE figures in the appendix (Figure 13). This is to be anticipated as inflation data are available at a higher frequency. As with GDP, the financial crisis weighs heavily on the forecast error and the post-crisis period is characterised by worse forecasting performance (Figure 13 in the Appendix, lower panel). Unlike GDP, though, the recent period (2014-2016) has been marked by significant and negative forecast errors so strong that even the most pessimistic forecasts turned out to be too optimistic: i.e. the min-max range of the forecasts is always in negative territory. This pattern has been reversed in the last two years when inflation picked up (2017-2018).

4 Optimality and rationality of Eurosystem/ECB forecasts

In this section we perform some tests on the series of the projection errors for GDP and inflation. Before moving on to the main parts of the analysis, it is worth emphasising some caveats. First, the analysis is conducted on a sample which is not long enough to derive concrete conclusions. In general, the test-statistics that are used below are valid asymptotically and although t-tests are used where possible - or small-sample adjustments are made - one cannot be entirely sure about the properties at short samples[7]. The second caveat relates to the accuracy of the tests used for evaluating *conditional* forecasts, and indeed when the underlying model that was used is not known. Although a lot of work has been done in the literature to correct for large models' parameter estimation error arising from nested models (for model-based forecasts) in performing these tests (West, 1996; West and McCracken, 1998; Clark and McCracken, 2001; McCracken,

[7] Similar analysis for the Federal Reserve projections are typically based on longer samples - see for example in Romer and Romer (2000) where the sample spans over 1965-1991. The BoE-IEO (2015) sample is similar to ours, 1997-2014.

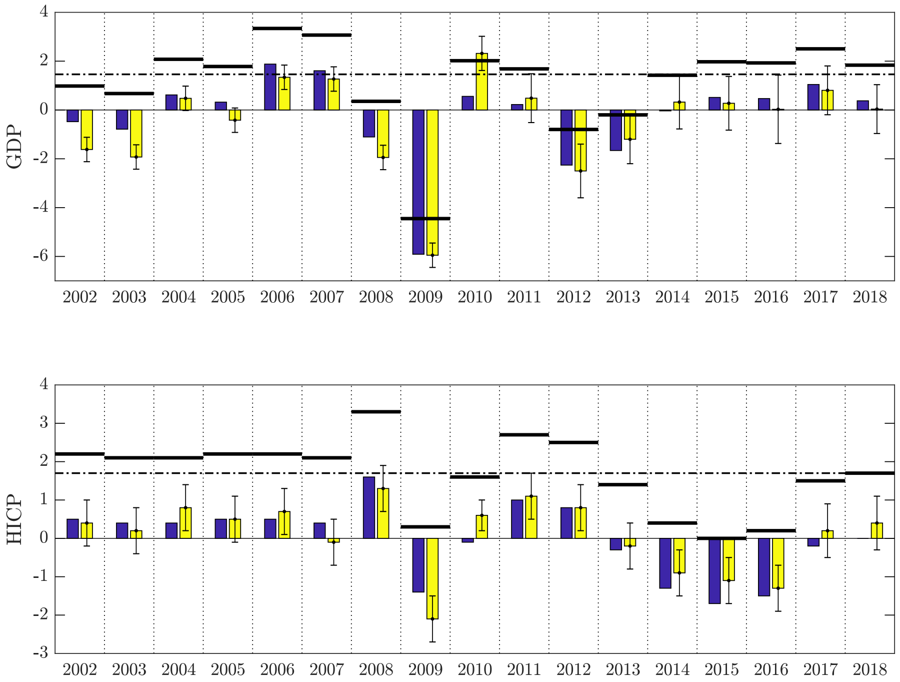

Figure 1: Projection errors cyclical pattern

Notes: Solid lines show calendar-year outcomes and dot-dash lines show the time series' long run mean using the latest vintage. Blue bars show actual deviations from the long run mean. Yellow bars show the June BMPE next calendar-year forecast errors and the error bands are the min-max forecast error range.

2007) - much less analysis exists on the asymptotic and small sample properties of conditional forecasts and the implications of errors in the conditioning variables[8]. To this end, to the extent that adjusted errors accurately represent the forecast error that would be made even if the actual path of the conditioning variables was known, the asymptotic properties of standard tests are valid. And this brings us to the final caveat of our approach. An implicit assumption behind the way adjusted forecast errors are computed is that the BMEs represent a good proxy of the main "(B)MPE model".

4.1 Unbiasedness

The test of unbiasedness is effectively a test on whether the mean of the forecast error is zero. If a forecast is unbiased, then, on average, there should not be any systematic over- or under-prediction and, on average, the mean of the projection error should not to be significantly different from zero. Hence, unbiasedness is a necessary property of optimal and rational forecasts. To test for unbiasedness we employ two different tests. First, we conduct a standard Mincer and

[8] See Faust and Wright (2008) and Clark and McCracken (2017).

Zarnowitz (MZ, 1969) regression-based test which is employed regularly (see for example Romer and Romer, 2000; El-Shagi et al., 2016):

$$y_t = c + \beta f_t + u_t \tag{7}$$

where y_t is the variable of interest as in the data (outcome) - using the conventional release of $t+5$ periods - and f_t the forecasted value of that variable as a regressor. We use heteroscedasticity and autocorrelation consistent covariance estimators (HAC) to account for autocorrelation in the forecast errors. Unbiasedness instructs that under the null hypothesis $H_0 : c = 0$ and $\beta = 1$.

Holden and Peel (HP, 1990) have shown that this is a sufficient but not necessary condition for unbiasedness and that MZ tends to over-reject. A necessary condition for unbiasedness is then the following HP test (see also Gavin and Mandal, 2003):

$$e_t = c + u_t \tag{8}$$

The forecast error is regressed on a constant and then we check - using HAC standard errors - whether this constant is equal to zero under the null hypothesis, i.e. $H_0 : c = 0$. In order to check the robustness of the results we conduct the t-test using two specifications of the HAC standard errors. Firstly, we use HAC standard errors as in Newey and West (1987) estimated using Bartlett kernel and a bandwidth defined by the theoretical autocorrelation properties of $h-$horizon forecast errors - that is $h-1$. The second specification employs Andrews (1991) data dependent automatic bandwidth selection method.

Results from the MZ tests are presented in Table 1 and the HP tests in Table 2 and graphically, for all horizons, in Figures 11 and 12 in the Appendix. Based on the combined results of both tests, Eurosystem/ECB forecasts for GDP and inflation forecasts are unbiased, except for GDP forecasts at long horizons.

Focusing first on the MZ tests, the joint null hypothesis of $c = 0$ and $\beta = 1$ for the "total error" (first two columns) is not rejected at 5% significance level for both GDP and inflation. Results are particularly strong for HICP inflation, where the joint null hypothesis is not rejected at any conventional level, at any forecast horizon and irrespective of looking at total or adjusted errors. For GDP, the null hypothesis is rejected at 10% level at short and long horizons but, interestingly, it is rejected strongly at 1% level when looking at adjusted errors.

Turning now to the HP tests, the constant in the regression 8 above tends to be negative

Table 1: Unbiasedness - MZ test

	GDP	HICP	GDP-adj.	HICP-adj.
		$h = 1$		
Constant (c)	-0.03	-0.03	-0.03	-0.03
P-value (c)	0.79	0.14	0.75	0.08*
Coefficient (β)	1.09	1.01	1.10	1.01
P-value (β)	0.13	0.25	0.06*	0.33
F-test p-value	0.06*	0.30	0.01*	0.12
		$h = 4$		
Constant (c)	-0.38	-0.45	-0.61	-0.36
P-value (c)	0.32	0.45	0.08*	0.09*
Coefficient (β)	1.00	1.28	1.38	1.17
P-value (β)	1.00	0.41	0.04**	0.16
F-test p-value	0.38	0.70	0.13	0.23
		$h = 8$		
Constant (c)	1.47	0.21	-2.52	0.23
P-value (c)	0.23	0.93	0.00***	0.68
Coefficient (β)	-0.29	0.90	2.30	0.76
P-value (β)	0.13	0.94	0.00***	0.40
F-test p-value	0.08*	0.98	0.00***	0.26

Results of the MZ regression - equation 7. P-values estimated using HAC (Bartlett) standard errors with bandwidth set according to Andrews (1991). F-test refers to the joint null $c = 0$, $\beta = 1$. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

for GDP growth for horizons beyond one quarter, implying on average some tendency to over-predict at $h > 1$. A similar, but much weaker, tendency for inflation is observed only after having adjusted the forecast error for the errors in the assumptions (we come to this later). In terms of statistical significance, however, there appears to be no systematic bias in the projections of HICP inflation at any forecast horizon (see also Figure 12) and thus, combined with the MZ results in Table 1, we can confidently conclude that HICP inflation forecasts are unbiased over the (full) sample we are looking at. On the other hand, the null hypothesis of the HP test for GDP at $h = 8$ is strongly rejected which, combined with the MZ results in Table 1, suggests a failure of the unbiasedness property of optimal forecasts. Figure 12 shows that the null hypothesis is rejected at 5% level for all $h \geq 6$. The bias remains when errors are adjusted for the errors in assumptions at $h = 8$. The results should nevertheless be interpreted with some care. Given that the GDP forecasts "pass" the MZ tests at 5% significance level, which is a sufficient condition for the lack of systematic bias, in principle we do not need to compute the HP tests and we could conclude outright that the GDP forecasts are unbiased (see Gavin and Mandal, 2003). Yet, the GDP forecasts only "weakly" pass the MZ test, as the null hypothesis is not rejected at 5% level but it is at 10% level at $h = 8$. At the same time, the strong rejection of the null hypothesis under the HP tests for the errors at $h = 8$ is something that we cannot simply ignore - especially as in some studies this is considered to be the standard test for unbiasedness (Clements et al., 2007; BoE-IEO, 2015). Overall, the weak rejection of the MZ null hypothesis and the strong rejection of the HP null hypothesis lead to the conclusion of a systematic bias in the long-horizon GDP forecasts.

Looking at results of similar studies, evidence on unbiasedness is mixed. The BoE-IEO (2015) finds that both the GDP and inflation forecasts of the BoE generally fail the MZ tests based on a sample similar to ours, but the HP tests do not provide significant evidence of a bias. The only exception is inflation at $h = 4$ for the overall sample at 10% level (they also show results excluding the financial crisis). Romer and Romer (2000) find that the Greenbook inflation forecasts - produced by staff at the Board of Governors of the Federal Reserve - and SPF inflation forecasts are "rational" - based on MZ type of tests - at almost all horizons but this cannot be said about the Blue Chip forecasts. Clements et al. (2007) also do not find any evidence of systematic bias in the Greenbook inflation and unemployment forecasts when employing regressions as in equation 8 above. On the other hand, Gavin and Mandal (2003) find that the forecasts of the members of the Federal Reserve Open Market Committee (FOMC) are unbiased for GDP but are strongly biased for inflation. Finally, Melander et al. (2007) find that the European Commission's GDP growth and inflation projections at both the EU and euro area level are unbiased, following a similar approach to the current study.

4.1.1 The difference in the results of total and adjusted errors

We close this section by taking a closer look into the difference between the total and adjusted errors. It is noticeable that the performance of the forecasts adjusted for the errors in assumptions relatively worsens across both tests; in so far that in both cases p-values generally move closer to the rejection region under the adjusted forecast errors (except for GDP at $h > 1$ under the HP tests). This might be hard to justify as one would expect a superior forecasting performance once the error in the conditioning assumption is removed. Indeed, the RMSEs of the adjusted errors are smaller than the total errors - see later in Table 6 - implying higher forecasting accuracy as expected. Nevertheless, statistically this implies a lower p-value (or equivalently a higher t-statistic), all else being equal. Therefore, to the extent that the mean error does not move closer to the value under the null - e.g. zero under the HP tests - then the t-statistic increases

Table 2: Unbiasedness - HP test

	GDP	HICP	GDP-adj.	HICP-adj.
	$h = 1$			
Bias	0.05	-0.01	0.06	-0.02
P-value (bw $h - 1$)	0.44	0.27	0.26	0.06*
P-value (bw Andrews)	0.60	0.27	0.45	0.06*
	$h = 4$			
Bias	-0.38	0.02	-0.19	-0.07
P-value (bw $h - 1$)	0.26	0.91	0.44	0.48
P-value (bw Andrews)	0.11	0.92	0.37	0.48
	$h = 8$			
Bias	-1.00	0.05	-0.58	-0.21
P-value (bw $h - 1$)	0.04**	0.85	0.13	0.31
P-value (bw Andrews)	0.00***	0.85	0.08*	0.17

Bias refers to the value of the constant in equation 8. The p-values are calculated using HAC (Bartlett) standard errors with bandwidth (bw) set to $h-1$ and according to Andrews (1991). *, **, *** indicate the null hypothesis is rejected at 10%, 5% and 1% significance level respectively.

making it more likely to reject.

In this regard, the relative worsening of the results for GDP adjusted errors in the MZ test are difficult to justify. The p-value of the MZ test is always lower for adjusted GDP errors, especially at medium-to-long forecasting horizons, stemming from both more negative constants and a β-coefficient that is higher than unity. One reason might be the failure of the implicit assumption stated above: that actually the BMEs are not a good proxy of the main "(B)MPE model". Indeed, in conducting forecasts a variety of models is used, as well as expert judgment, and the BMEs are only one member of the toolkit the Eurosystem/ECB forecaster possesses in doing the forecast. It might indeed be that the BMEs underestimate the impact of the errors in assumptions. That would still push the forecast in the right direction, implying a decrease in the absolute or squared forecast error, but not enough such that to eliminate the bias. On the other hand, it could be that the representative Eurosystem/ECB forecaster was persistently more optimistic about the long-run GDP than what was suggested by the models at her disposal; including the BMEs. For that, it would be interesting to check how this over-optimism evolved over time and what has been the impact of the crisis - which we do later on.

For the case of inflation, the relative worsening of the results seems to be related primarily to the higher forecasting accuracy, except possibly the HICP inflation forecasts at long horizons $h = 8$. In the case of the MZ tests, all values move closer to the theoretical ones with the only exception the β coefficient under $h = 8$. Similarly, at $h = 8$ under the HP tests the constant is markedly different from zero when compared to the total error. On the other hand, the relative deterioration of the results at short horizons (the HP test rejects the null at 10% level) could possibly be attributed to the fact that different models are used to forecast inflation at horizons below one year. In particular, HICP inflation and components up to one-year horizon are produced through the NIPE - see Section 2 and ECB (2016). Indeed, to the extent that the elasticities of the BMEs are significantly different than those implied by the NIPE models the appropriateness of the adjustment performed with the BMEs is somewhat questionable. It is, however, very difficult to cross-check the validity of this premise which we leave for further research. Regarding HICP forecasts at longer horizons, the deviation from the optimal theoretical value could also be related to the rightful use of the BMEs as described for GDP above. However, for the case of inflation this has to be also seen against a "very" unbiased overall forecast. That is, the constant under the HP tests is very close to zero and arguably it could hardly be improved further. The main "(B)MPE model" and/or the representative Eurosystem/ECB forecaster appear to be unbiased. Thereby, adjusting the errors in assumptions does improve forecasting accuracy but does not necessarily translate in "less bias".

An avenue to evaluate the potential of the reasons outlined in this section to explain this à priori unexpected result would be to employ alternative model(s) to correct the forecast for the errors in assumptions and compare the results. We leave this for future research.

4.1.2 Pooled approach

The approach so far has concentrated on testing for bias for a specific variable and at a specific horizon. In this section we test whether forecasts are collectively unbiased by pooling together information from forecasts at all horizons, following the techniques in Clements et al. (2007) and Ager et al. (2009). Indeed, as forecasts of rational agents (with squared loss functions) should be unbiased at all horizons, our previous finding of bias in GDP forecasts at long, but not at short, horizons is somewhat difficult to interpret (Clements et al., 2007).

We conduct two type of tests: one in which we impose a common bias c, across all horizons, and subsequently test whether this is significantly different from zero. This tests whether there is a systematic common bias, on average, across all horizons. This amounts to equation (7) in Clements et al. (2007):

$$e = i_{TH}c + v \tag{9}$$

where the $TH \times 1$ vector e contains the stacked error terms $e = Y - F$ with outcomes y_t stacked in vector $Y = (y_1, y_2, ..., y_T)' \otimes i_H$ and forecasts in $F' = (f_{1H}, ..., f_{11}, ..., f_{TH}, ..., f_{T1})$. i_H and i_{TH} are unit vectors of dimension $H \times 1$ and $TH \times 1$ respectively. v is the stacked sum of the aggregate macroeconomic (u_{th}) and idiosyncratic (ϵ_{th}) shocks that occurred between $t-h$ and t[9]. From (9) it is clear that the bias c is restricted to be the same across all horizons (see Clements et al., 2007).

In a second test, we allow for horizon-specific bias and subsequently we test whether these horizon-specific biases are *jointly* equal to zero. As in equation (8) in Clements et al. (2007) we have:

$$e = (i_T \otimes I_H)C_H + v \tag{10}$$

where the $H \times 1$ vector C_H contains the separate bias for each horizon: $C_H = (c_H, ..., c_1)$. The explanatory variables of the regressions are contained in matrix $(i_T \otimes I_H)$ of dimension $TH \times H$, where I_H is the identity matrix of order H.

We compute the Wald statistic distributed as χ^2 with h degrees of freedom by using a subset $C_\hbar$ of vector C_H for $h = 1, ..., \hbar, ..., H = 8$. As Ager et al. (2009) explain, this sequential approach can help assess at which forecast horizon the bias becomes significant. More formally we compute the Wald statistic as in equation (5) in Ager et al. (2009)[10]:

$$W_h = (\hat{C}_\hbar)[var(\hat{C}_\hbar)]^{-1}(\hat{C}_\hbar) \tag{11}$$

We perform tests both with and without idiosyncratic shocks. However, as it turns out idiosyncratic shocks tend to have zero variance, and hence do not affect the results, in all except the case of GDP adjusted errors. Therefore we focus on results from common, macroeconomic shocks which capture forecast errors due to fundamental economic shocks. The results are presented in Table 3 and graphically in Figure 12. Overall, when assuming a common bias

[9] As in Clements et al. (2007) we assume the forecasts are produced in period t-1 for period t for $h = 1$. This is also aligned with the definition of the error at $h = 1$ in the context of the paper. Although this is actually a nowcast, only a limited set of information is available for the current quarter when the forecast is made.

[10] Following Greene (2003), the Wald statistic that is χ^2 distributed is transformed to the corresponding F-statistic by dividing the value of the Wald statistic by its degrees of freedom.

Table 3: Unbiasedness - Pooled approach

	GDP	HICP	GDP-adj.	HICP-adj.
Common bias	-0.45	0.02	-0.23	-0.10
Common bias p-value	0.33	0.95	0.48	0.51
Horizon-specific bias: F-test p-value	0.00***	0.00***	0.00***	0.00***
Horizon at which the forecast becomes significantly biased - Figure 12	2	7	2	2

Common bias refers to the constant in equation 9. The p-value is estimated using OLS with no idiosyncratic shocks as in Clements et al. (2007). F-test p-value refers to the joint null hypothesis of horizon-specific bias equal to zero. Tests are conducted after pooling all horizons $1 \sim 8$ without idiosyncratic shocks, as in Ager et al. (2009).*, **, *** indicate the null hypothesis is rejected at 10%, 5% and 1% significance level respectively.

across all horizons, we cannot reject the null hypothesis that this common bias is equal to zero for both GDP and HICP, for "total" and "adjusted" errors. The null hypothesis is not rejected also when we implement the test sequentially - i.e. in sequential tests of a zero common bias up to $h = 2, h = 3, ..., h = 8$ (see Figure 12). However, once we allow for horizon-specific bias the null hypothesis of zero bias at all $h = 1, ..., 8$ is strongly rejected, for both variables and for both types of forecast errors (i.e. we compute W_8 statistic above). The latter is the first strong evidence of a bias in the Eurosystem/ECB forecasts that we document in this study, for both GDP and HICP. Furthermore, the failure to reject the null hypothesis of a common bias suggests that the unbiasedness hypothesis fails at some, or after a certain forecasting horizon. We test this premise by conducting a series of sequential F-tests at different horizons $h = 1, ..., 8$ as in Ager et al. (2009). In Table 3 we present the horizon at which forecasts become significantly biased - graphically the results of these tests are shown in Figure 12 in the appendix. For GDP, the answer is rather discouraging: GDP forecasts are biased for any $h > 1$ when tested with this method. For HICP inflation, the picture is more positive for the Eurosystem/ECB forecaster as the bias "kicks-in" only at long horizons, in particular at $h = 7$. Nevertheless, we again have the unexpected result of stronger bias for adjusted errors, which appears already as of $h = 2$.

Whether we assume a common bias or we allow for a horizon-specific bias is of crucial importance to our results. The sensitivity to this assumption is documented also in Clements et al. (2007) for the FED's GDP, inflation and unemployment forecasts as well as in Ager et al. (2009) for the Consensus Forecasts for GDP and inflation. Similarly to our case, Clements et al. (2007) find that the null hypothesis of a common zero bias is not rejected with high p-values, whereas it is strongly rejected at 1% level when horizon-specific bias is allowed. Ager et al. (2009) find similar results for GDP growth and inflation Consensus Forecasts.

4.1.3 Time-varying bias

So far our analysis has concentrated in evaluating the bias of the Eurosystem/ECB forecasts over the full sample. Nevertheless, we have shown in Figure 1 that the forecast performance has been changing over different parts of the business cycle. For example, it is well known that over the period of "low inflation" 2012 - 2016 there had been an episode of persistent over-predictions of inflation, a fact that has been mentioned also in speeches and other communication by ECB Board members (Constâncio 2015, also Figure 10). Bobeica and Jarociński (2017) and Cicarelli and Osbat (2017) argue that a reason behind the missing disinflation was the use of models that were too restrictive and did not sufficiently allow for domestic as well as external factors in driving inflation. Indeed they show that global factors were key in understanding inflation dynamics during that period. On the other hand, ECB (2013) finds that GDP was over-estimated and HICP inflation under-estimated over 2000-2012. In this section we aim to perform some tests on the time variation of the forecast bias[11].

In particular, we perform the same tests as in the previous section on rolling windows of 25 quarters as El-Shagi et al. (2016). The previous section has documented strong evidence of bias in the GDP projections at long horizons and no bias in inflation except under the pooled approach. In Figures 2 and 3 we observe that indeed the GDP forecast at $h = 8$ has been significantly biased across almost the whole sample[12]. The constant in equation 7 is significantly above zero almost at all times, and similarly the β coefficient is significantly below one, with the exception of the period between the two crises. As a consequence, the rolling-window F-test for the joint null hypothesis of $c = 0$ and $\beta = 1$ is rejected along the whole sample for $h = 8$, except from the inter-crises period (Figure 2, lower panel). At shorter horizons ($h = 4$) the results are more in line with the theory of optimal forecasts, although we observe some deviations from the optimal values before the financial crisis. Looking at rolling HP tests (Figure 3, upper panel), the evidence of a bias is less strong, with the zero-line being always within the confidence bands at $h = 4$ and outside occasionally before the financial crisis and after the sovereign-debt crisis for $h = 8$. The rolling F-tests under the pooled approach reject the null hypothesis of zero horizon-specific bias (pooled across horizons) for both $h = 4$ and $h = 8$ in the early and latest parts of the sample (Figure 3, lower panel). Overall, the time-varying bias tests for GDP suggest

[11] We do not look at adjusted errors in this case. This allows to use the full sample of forecast errors that starts in 1999Q1.

[12] For $h = 1$ both GDP and inflation forecasts perform well so we do not discuss those in this part.

that biased forecasts were concentrated in the early and latest parts of the sample. Thus, the evidence of a negative bias in the GDP (B)MPE forecasts (over-prediction) across the whole sample - especially at long-horizons - discussed previously appears to be primarily a result of biased forecasts over these periods.

In terms of inflation, the rolling window tests indicate the presence of statistically significant bias (under-prediction) in the first few years of the existence of the Eurosystem/ECB staff forecasts and up to the crisis, but no significant bias since then. The MZ rolling window tests provide evidence of significant departure of c and β from their optimal values for $h = 4, 8$ on several occasions - primarily in the early and latest parts of the sample for $h = 4$ and since the financial crisis for $h = 8$, except for the last few observations (see the rolling F-tests in the lower panel of Figure 2). These are indeed concentrated over the periods that the Eurosystem/ECB staff are known to have been making forecast errors persistently in a single direction - i.e. to under-predict inflation in the pre-crisis period and to over-predict it in the more recent low-inflation period. Nevertheless, rolling-window HP tests in Figure 3, upper panel, show a more clear time-varying pattern: they indicate that inflation forecasts have been biased towards under-prediction in the pre-crisis period, but do not provide enough evidence of a statistically significant bias since the financial crisis. Actually, by looking only at the HP regressions, it appears that the inflation forecasts' bias at $h = 4, 8$ has been steadily decreasing over time - from a persistent under-prediction of inflation (significant) to a persistent over-prediction (non-significant). These results are not at odds with the well documented persistent negative forecast errors on inflation over the recent years, which have attracted the focus of policy makers and the financial press. Indeed, a persistent negative bias (over-prediction) towards the end of the sample is visible in Figure 3 (upper panel, for $h = 4, 8$); as well as by the spike in the β coefficient in the MZ regressions over the recent years bringing it significantly above unity (Figure 2, middle panel, $h = 4$). What this analysis shows is that it cannot be concluded with certainty that this bias is significant in a statistical sense, given the sample and the data in hand. That does not imply that these forecast errors are fully justified and could not have been avoided, and it does not imply that the policy decisions taken over this period based on this, but as well as other information, would not have been different. It is thus mostly a statistical claim. Finally, the evidence of horizon-specific bias under the pooled approach discussed above seems to be also mostly related to the bias in the early parts of the sample, as the rolling F-tests reject much less frequently the null hypothesis of zero horizon-specific bias (pooled across horizons) since the financial crisis (Figure 3, lower panel, for $h = 4, 8$).

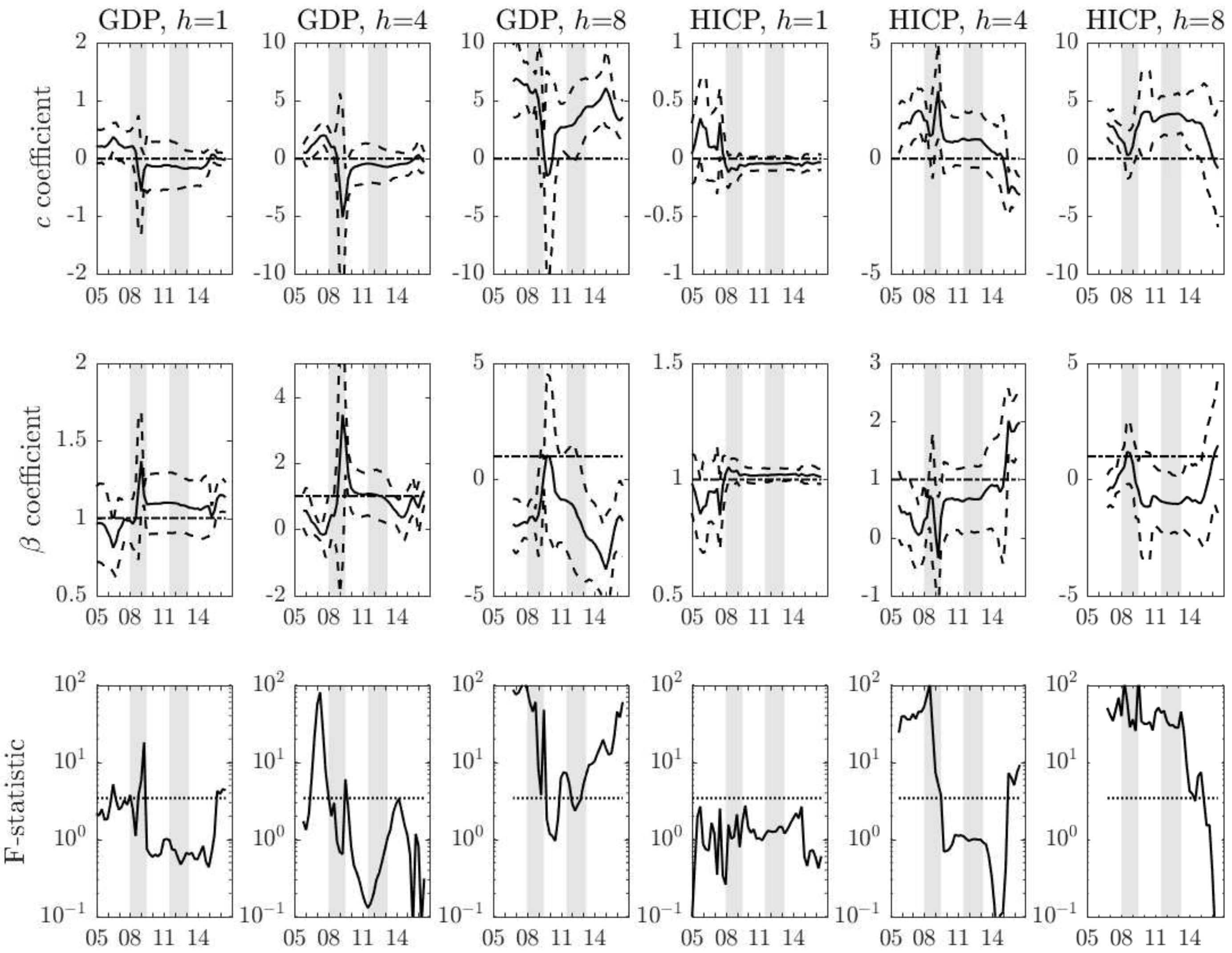

Figure 2: Rolling-window estimates of bias - MZ test

Notes: First and second panel: c and β estimates of the MZ equation 7 for h =1, 4 and 8 and their 95 % confidence intervals. Third panel: F-statistic for the joint hypothesis that $c = 0$ and $\beta = 1$. Dashed lines: F-test critical value at 5% significance level. HAC (Bartlett) standard errors with bandwidth set according to Andrews (1991). The window size is 25 observations. Y-axis is linear unless otherwise specified as logarithmic.

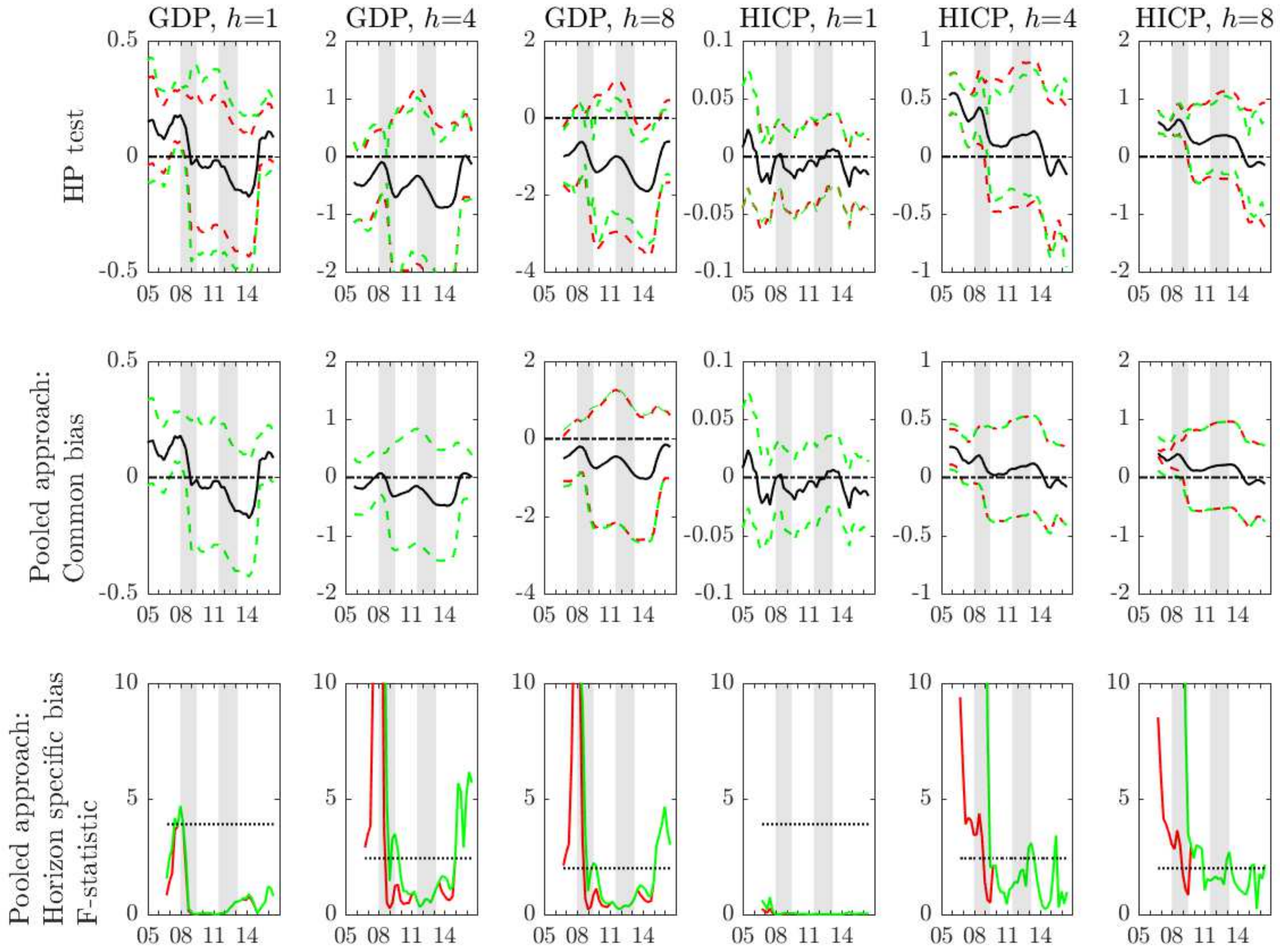

Figure 3: Rolling-window estimates of bias - HP and pooled-approach tests

Notes: First panel: Constant in the HP equation 8 and 95 % confidence intervals (red: bandwidth set to $h-1$, green: bandwidth set according to Andrews 1991). Second panel: Clements et al. (2007) pooled over all horizons common bias and 95 % confidence intervals (red: with idiosyncratic shocks, green: without idiosyncratic shocks). Third panel: Ager et al. (2009) F-statistic for the joint hypothesis of zero horizon-specific bias (red: with idiosyncratic, green: without idiosyncratic shocks). Dashed lines: F-test critical value at 5% significance level. The window size is 25 observations.

4.2 Efficiency

Forecasts are efficient if they are unpredictable. Statistically, this could be tested in an extended form of the MZ regression above:

$$y_t = c + \beta f_t + \gamma X_t + u_t \tag{12}$$

where X_t is any variable in the forecasters' information set Ω_t - i.e. any variable that was available at the time that the forecast was made - $X_t \in \Omega_t$. Efficiency would thus require that this variable be orthogonal to the forecast error - that is $\gamma = 0$, $\forall\, X_t \in \Omega_t$ (Keane and Runkle, 1994). Weak efficiency, as in El-Shagi et al. (2016) could be tested if the past forecast error has no explanatory power for the dependent variable - i.e. $\gamma = 0$ and $X_t = e_{t-1,1}$.

We follow a slightly different approach by regressing the forecast error on all possible combinations of past errors and outcomes as in Gavin and Mandal (2003).

$$e_t = c + \gamma X_t + u_t \tag{13}$$

where X_t is lagged GDP/HICP error or outcome[13]. We use one explanatory variable at a time such that there are 4 possible combinations of the above equation for each error in GDP and inflation. If Eurosystem/ECB staff forecasts are informationally efficient, they would take into account all recent GDP and inflation data as well as the previous forecast errors. Therefore, past GDP/HICP inflation errors and data should be orthogonal to current forecast errors and the coefficient γ in the above regression should not be significantly different from zero. Results are presented in Table 4 (p-values of parameter γ in (13)).

In general, HICP forecasts are informationally efficient but that cannot be said for GDP forecasts. From all the 12 possible combinations we look at for each variable's forecast error, the null hypothesis of informational efficiency in the HICP inflation forecasts is rejected only once and indeed only at 10% level (case against the GDP error at $h = 4$). For GDP, informational efficiency fails at 8 out of the 12 cases; one of those at 10% level. For GDP forecasts up to one year it appears that inflation outcomes had not been properly taken into account suggesting

[13] It is noted that "outcome" in this case would refer to the first release of GDP/inflation. The forecast error used as an explanatory variable X_t is also computed using the first release. It is understood that the first estimate of GDP for each quarter might actually not have been made available until late in the preparation of a forecast. We nevertheless abstract from these considerations and assume that the first estimate was available. The results are robust to using the GDP at $t - 2$ periods before the actual date.

Table 4: Efficiency tests

	GDP outcome	GDP error	HICP outcome	HICP error
		$h = 1$		
GDP error	0.38	0.06*	0.00***	0.93
HICP error	0.12	0.30	0.94	0.77
		$h = 4$		
GDP error	0.00***	0.43	0.01**	0.01*
HICP error	0.25	0.08*	0.94	0.59
		$h = 8$		
GDP error	0.00***	0.00***	0.23	0.00***
HICP error	0.83	0.28	0.64	0.53

P-values for the null hypothesis $\gamma = 0$ in equation 13 estimated with HAC (Bartlett) and bandwidth set according to Andrews (1991). The columns are the choice of regressor X_t in equation 13. They refer to each (B)MPE's previous quarter observed data, using the first release, and the respective $h = 1$ error estimated using the same outcome. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

some inconsistency between the GDP and inflation forecasts. This could also relate to the fact that inflation forecasts up to one year are performed under a different exercise, the NIPE. Nevertheless, GDP forecasts at $h = 4, 8$ fail to take into account information from own past GDP data.

Another test for weak efficiency is the autocorrelation in the forecast errors. Assuming a MSE loss function, the autocorrelation of forecast errors should be up to $h - 1$ (Elliot and Timmermann, 2016). The autocorrelation properties of the euro area GDP growth and inflation forecast errors are tested by employing the Cumby and Huizinga (CH, 1992) and the Ljung and Box (LB, 1978) tests.

The CH test checks whether a time series exhibits serial correlation of a given order, against the hypothesis of autocorrelation of some higher order. In the context of our analysis we test for the possibility of a serial correlation in the forecast error going up to order $h+4$, instead of only up to $h-1$ as implied by the theory. The CH's test hypothesis testing is:

$$\begin{aligned} &CH\ test \\ &H_0: \text{ Forecast errors are MA(h-1), for } h \geq 1 \\ &H_1: \text{ Forecast errors' autocorrelations of} \\ &\ \text{order } \tau \text{ are non-zero, for } h \leq \tau \leq h+4. \end{aligned} \tag{14}$$

Rejecting the null hypothesis would provide evidence of serial correlation in the forecast error beyond what is implied by the theory, potentially up to order $h+4$.

The LB's test hypothesis testing is:

$$\begin{aligned}
&\textit{LB test}\\
&H_0: \text{ Forecast errors' autocorrelations up to order } \tau \leq h-1\\
&\text{ are zero, for } h \geq 2\\
&H_1: \text{ Otherwise}
\end{aligned} \tag{15}$$

Performing both tests protects against type I and type II errors. Results are provided in Table 5. The autocorrelation of the projection error of euro area GDP growth and inflation are in line with the theory of optimal forecasts, as the null hypothesis of the CH test is *not* rejected and the null of the LB test *is* rejected. The only exception is GDP growth projection errors at $h=1$ which appear to be autocorrelated, in contrast with the theoretical implication of zero autocorrelation. This result applies to both total forecast errors and forecast errors adjusted for errors in the conditioning assumptions. In Figure 4 we provide the tests' results at all horizons and also show the correlogram of the total errors at $h=1,4,8$. Indeed, we see that weak efficiency fails only for $h=1$ errors of GDP. Looking at the correlogram of this error, we see that this failure arises because of significant autocorrelation of order one. Apart from that, autocorrelations of higher order are indeed not statistically significant for both GDP and HICP errors - although there are some marginal cases.

Gavin and Mandal (2003) perform similar tests as in Table 4 for the forecasts of the Federal Open Market Committee (FOMC). They find that FOMC forecasts are informationally efficient and that out of the 48 cases examined, they could reject the orthogonality in only three - 2/24 for GDP and 1/24 for inflation. Thus, the members of the FOMC GDP forecasts appear to make better use of available data than the Eurosystem/ECB staff forecasts do, although they somehow appear to be more biased in their inflation forecasts than Eurosystem/ECB staff is. Alessi et al. (2014) show that the (B)MPE and the Federal Reserve Bank of New York (FRBNY) forecasts could have been improved - especially during the financial crisis - had they taken into account

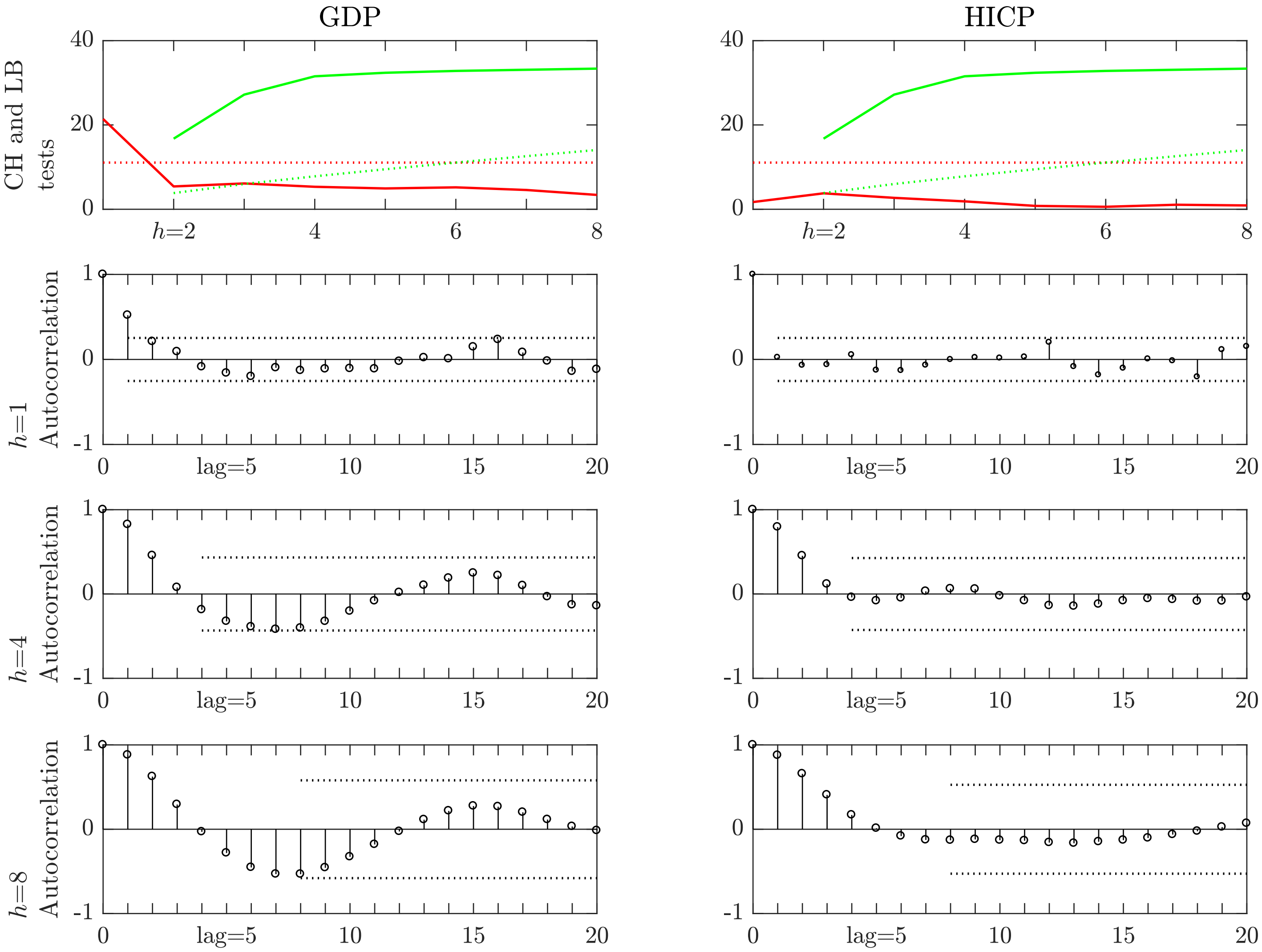

Figure 4: Autocorrelation tests and correlograms of the forecast errors
Notes: First panel: CH test (Cumby and Huizinga 1992, red line) and LB test (Ljung and Box 1978, green line) and corresponding critical values at 5% significance level in dashed-lines with same colours. Remaining panels: correlograms of GDP and HICP total forecast errors at different forecasting horizons. Dotted lines show the 5% confidence interval under the null hypothesis of an autocorrelation of order $h-1$ - corresponding to $1.96 \times SE$. The estimated standard error (SE) of the autocorrelation at lag k, (r_k), $k > q$ is: $SE(r_k) = \sqrt{\frac{1}{T}(1+2\sum_{j=1}^{q} r_j^2)}$

Table 5: Autocorrelation tests

	GDP	HICP	GDP-adj.	HICP-adj.
		$h = 1$		
CH p-value	0.00***	0.89	0.00***	0.16
LB p-value	-	-	-	-
		$h = 4$		
CH p-value	0.38	0.87	0.53	0.62
LB p-value	0.00***	0.00***	0.00***	0.00***
		$h = 8$		
CH p-value	0.64	0.97	0.89	0.65
LB p-value	0.00***	0.00***	0.00***	0.00***

CH: Cumby and Huizinga (1992). LB: Ljung and Box (1978) tests. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

high-frequency financial data. In BoE-IEO (2015, Section 3.3) a similar test is conducted and, similarly to our case, the results are less encouraging for GDP than for inflation - excluding crisis periods. In particular, it is found that the GDP forecast error at $h = 8$ is related to past forecast errors and the GDP forecast error at $h = 1, 8$ is related to past GDP outcomes - thus in 6 possible combinations checked in a fashion similar to above (GDP error and GDP outcome for $h = 1, 4, 8$) efficiency is rejected three times against one out of six for inflation (relation to past inflation outcomes at $h = 8$). Melander et al. (2007) find that the EC forecasts of GDP and inflation are weakly efficient at the euro area and EU aggregate level. Contrary to the current results for the Eurosystem/ECB, the EC inflation current-year forecasts for the euro area and the EU do not adequately reflect available information on past inflation.

4.3 Standard deviation - RMSE

Table 6 depicts the standard deviation of the forecast errors, the Root Mean Squared Errors (RMSEs) and the scaled RMSEs with their standard errors. The standard deviation of the forecast error increases with the projection horizon in accordance with the theory[14]. RMSEs and scaled RMSEs are also increasing, which are defined as:

$$RMSE = \frac{1}{T}\sum_{i=1}^{T} {e_t}^2 \qquad (16)$$

[14] See Patton and Timmernann (2012) for a formal test on the monotonicity of MSE over horizons.

$$Scaled\,RMSE = \frac{1}{T}\sum_{i=1}^{T}\left(\frac{e_t}{\sigma_y}\right)^2 \tag{17}$$

The scaled RMSE is the RMSE scaled by the variance of the underlying variable, underscoring the difficulty in predicting more volatile variables. Graphs of the RMSEs at all horizons are provided in the Appendix (Figure 13). Forecast accuracy tends to lessen as the forecast horizon increased, in line with the property of optimal forecasts. As the forecast horizon increases, available data will provide a weaker signal about the likely path of a given variable. Moreover, there is greater scope for unforeseen shocks to occur as the forecast horizon lengthens. Furthermore, the 95% confidence intervals for scaled RMSEs at longer horizons do not overlap with the confidence intervals for scaled RMSEs at short horizons - mostly in the case of inflation than for GDP. This means we can be reasonably confident that the apparent differences in accuracy between short and longer-term forecasts are statistically different.

The RMSE can be decomposed into its variance and bias component as in the following equation:

$$RMSE = \sqrt{bias^2 + variance} \tag{18}$$

In Table 6, the difference between the RMSE and the standard deviation of the errors is attributed to the bias through the RMSE decomposition. This difference is most significant for GDP forecast errors at $h = 8$; indeed where we document the existence of a persistent bias in Section 4.1. Thus, avoiding systematic errors in one direction has two benefits: eliminating the bias and increasing forecasting accuracy through a lower RMSE. Finally, two remarks: (i) the accuracy of the forecasts improves - i.e. the RMSE decreases - once the errors in conditional assumptions are accounted for and (ii) inflation forecasts are more precise than GDP forecasts.

Figure 13 provides also estimates of the RMSEs before, after and excluding the financial crisis. This major economic event has a clear profound impact on forecast accuracy[15]. RMSEs excluding the crisis are always inferior to those that include the crisis and forecasting accuracy of both GDP and inflation was higher in the pre- than in the post-crisis period. This is related to the fact that the post-crisis period includes another important economic episode in the euro area - the sovereign-debt crisis[16].

[15] Indeed, BoE-IEO (2015) exclude the financial crisis from the main sample of the analysis.

[16] It should be noted however that the sample size of the sub-periods is not the same.

Table 6: Standard deviation - RMSE

	GDP	HICP	GDP-adj.	HICP-adj.
		$h = 1$		
St. deviation	0.49	0.08	0.39	0.07
RMSE	0.48	0.08	0.40	0.07
Sc. RMSE	0.28	0.08	0.23	0.07
		$h = 4$		
St. deviation	1.55	0.82	1.13	0.45
RMSE	1.58	0.81	1.14	0.45
Sc. RMSE	0.92	0.79	0.66	0.44
		$h = 8$		
St. deviation	1.95	1.04	1.41	0.78
RMSE	2.17	1.03	1.51	0.80
Sc. RMSE	1.25	1.01	0.87	0.78

5 Other tests of forecasting accuracy and performance

5.1 Normality

We examine the forecast errors on their normality, i.e. whether their skewness and kurtosis statistics resemble those of a normal distribution. Historical forecast errors are used by many institutions to provide ranges of uncertainty over their forecasts - see Tulip and Wallace (2012) for an overview of major central banks. Eurosystem/ECB ranges are estimated using the mean absolute error after removing statistical outliers - see ECB(2009). For this reason the distributional properties of forecast errors have attracted some attention in the literature. Harvey and Newbold (2003) in examining the Survey of Professional Forecasts (SPF) dataset for US GDP growth and CPI inflation forecast errors find evidence of skewed and leptokurtic distributions. Reischneider and Tulip (2007), under the caveats of using Jarque and Bera (JB, 1987) test on serially correlated data, overall find evidence of Fed's GDP growth forecast errors being normal but not for CPI inflation (see also Reifschneider and Tulip, (2018) for the impact of the crisis on these results).

The presence of serial correlation in forecast errors for $h > 1$ would however deem the results of many of the standard tests for normality - e.g. the Jarque and Bera (JB, 1987) not valid. The literature covering normality tests for serially correlated data among others includes Bontemps and Meddahi (2005), Lobato and Velasco (LV, 2004) and Bai and Ng (BN, 2005). In our analysis we resort to the latter two which test the same null as JB but use consistent estimators of the variance of skewness and excess kurtosis. In more detail BN rely on kernel estimators, while LV use sample estimates of the asymptotic variances which do not involve any kernel smoothing methods. BN propose also separate tests for the skewness and kurtosis of the data that we include in our analysis to have a better understanding of the forecast errors' distributional properties with respect to the symmetry and their tails. BN suggest separate tests for skewness (named π_3), kurtosis (named π_4) and normality defined as: $\pi_{34} = \pi_3^2 + \pi_4^2 \xrightarrow{d} \chi_2^2$. Performing these tests allows to capture whether the failure of normality, if so, results from skewed errors or fat tails or both. Formally, the null hypothesis of our class of tests on normality is that skewness and excess kurtosis are jointly zero.

Results of the normality tests are presented in Table 7. Overall, forecast errors of GDP growth and inflation appear to be negatively skewed, implying, on average, a higher probability of over-predicting these variables. Moreover, the kurtosis of GDP growth projection errors appear to be rather far from 3, contrary to the HICP inflation projection errors. In terms of statistical tests, overall, there is evidence of normality in HICP inflation errors but significant departures from normality are detected for GDP growth errors[17]. This is the reverse of the results found by Reischneider and Tulip (2007) on Fed's forecast errors. In the case of HICP inflation, all normality tests p-values' agree on evidence of normality - i.e. failure to reject the null at conventional levels. For GDP forecast errors at $h = 1$ evidence in favour of normality are mixed: there is disagreement in two cases between LV and BN for both total and adjusted errors - the LV test strongly rejects the null hypothesis of normality whereas the BN tests fails to reject. This could be related to the documented low power of BN test in simulation studies with small samples - see Bai and Ng (2005). However, there is strong evidence against normally distributed GDP forecast errors at $h = 4, 8$ as all tests agree on this conclusion. Looking closer into the results, this is due to strong evidence of leptokurtic distributions - i.e. "fat tails" - both in total and adjusted errors (BN-π_4 test). BN skewness test indicates symmetry of all distributions of the errors at 5% significance level. As the Section 4.3 has shown the largest forecast errors were naturally committed during the financial crisis (see also Figure 1), this episode might be a main reason behind the apparent fat tails of the GDP projection errors' distribution.

We test this hypothesis by conducting these tests across forecast errors of GDP and inflation at all horizons, with and without the financial crisis in the sample. Results are presented graph-

[17] Interestingly, the results from the JB test are in line with its counterparts even though autocorrelation in the forecast errors is not explicitly accounted for.

Table 7: Normality tests

	GDP	HICP	GDP-adj.	HICP-adj.
		$h = 1$		
Skewness	-2.53	-0.50	-1.69	-0.39
BN-π_3 p-value	0.24	0.07*	0.26	0.11
Kurtosis	12.43	2.89	8.88	2.65
BN-π_4 p-value	0.13	0.81	0.15	0.44
JB p-value	0.00***	0.28	0.00***	0.41
LV p-value	0.00***	0.26	0.00***	0.38
BN-π_{34} p-value	0.16	0.18	0.18	0.21
		$h = 4$		
Skewness	-1.90	-0.73	-2.25	-0.46
BN-π_3 p-value	0.20	0.22	0.21	0.27
Kurtosis	7.82	3.44	9.55	3.11
BN-π_4 p-value	0.03**	0.61	0.08*	0.85
JB p-value	0.00***	0.06*	0.00***	0.34
LV p-value	0.00***	0.23	0.00***	0.56
BN-π_{34} p-value	0.04**	0.41	0.10*	0.53
		$h = 8$		
Skewness	-1.61	-0.30	-1.36	0.01
BN-π_3 p-value	0.18	0.55	0.15	0.97
Kurtosis	5.68	2.05	4.93	2.46
BN-π_4 p-value	0.01***	0.33	0.10*	0.55
JB p-value	0.00***	0.27	0.00***	0.76
LV p-value	0.00***	0.61	0.01***	0.92
BN-π_{34} p-value	0.01***	0.52	0.09*	0.83

Normality tests: JB: Jarque and Bera (1987), LB: Lobato and Velasco (2004), BN (π_{34}): Bai and Ng (2005). Skewness (π_3) and kurtosis (π_4) p-values refer to two-sided tests under the null of zero and three respectively. P-values for all BN test statistics use HAC (Bartlett) standard errors with bandwidth set according to Andrews (1991). *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

Table 8: Contingency table

	$\Delta F > 0$	$\Delta F \leq 0$	Total
$\Delta Y > 0$	$N_{11} =$ $N(\Delta Y > 0, \Delta F > 0)$	$N_{12} =$ $N(\Delta Y > 0, \Delta F \leq 0)$	$N_{1.} =$ $N(\Delta Y > 0)$
$\Delta Y \leq 0$	$N_{21} =$ $N(\Delta Y \leq 0, \Delta F > 0)$	$N_{22} =$ $N(\Delta Y \leq 0, \Delta F \leq 0)$	$N_{2.} =$ $N(\Delta Y \leq 0)$
Total	$N_{.1} = N(\Delta F > 0)$	$N_{.2} = N(\Delta F \leq 0)$	N

ΔY represents the change in the underlying variable and ΔF the change in the forecast. The input in the cells is the number of times the condition is satisfied (i.e. upper-left corner when both $\Delta Y > 0$ and $\Delta F > 0$).

ically in Figure 5. The results improve but some departures from normality of GDP forecast errors remain. First, the green lines (excluding the crisis) in the panels of the JB, LV and BN tests are most often below the straight red lines - especially for the case of GDP - implying less evidence against the null hypothesis. In terms of the statistics, for GDP forecast errors we see that the LV test always rejects the null hypothesis of normality at all horizons, as in Table 7, whereas excluding the crisis period implies normality for forecast errors at horizons up to $h = 4$. Interestingly, over long forecasting horizons the LV test results of the full sample and the excluding the crisis sample are very close, which could be interpreted that even without the impact of the large errors made during the crisis, the forecast errors remain large enough to distort the normal distribution by fat tails. Contrary to the LV test, the BN test supports normality of GDP projection errors at all horizons when the impact of the crisis is accounted for. Nevertheless, this test has the lowest power of the three (Psaradakis and Vàvra, 2018). For inflation, we observe only one single departure from normality at $h = 3$ under the LV test which is “restored” once we exclude the financial crisis. Finally, for GDP we see that the impact of financial crisis is more important than the impact of assumptions over the full sample.

5.2 Directional accuracy

Tests for the directional accuracy check whether the change in the projected variable follows that of the outcome. That is, these tests focus on whether the forecaster has correctly predicted the direction - i.e. whether there is a pick-up or slowdown - in the year-on-year growth rate of the variable of interest, rather than checking the extent of over/under prediction. This is of course very important and for policy purposes a correctly projected path (e.g. inflation is increasing) can be more relevant than the value of the projected variable itself.

The various different combinations are often represented in the form of a 2 × 2 contingency

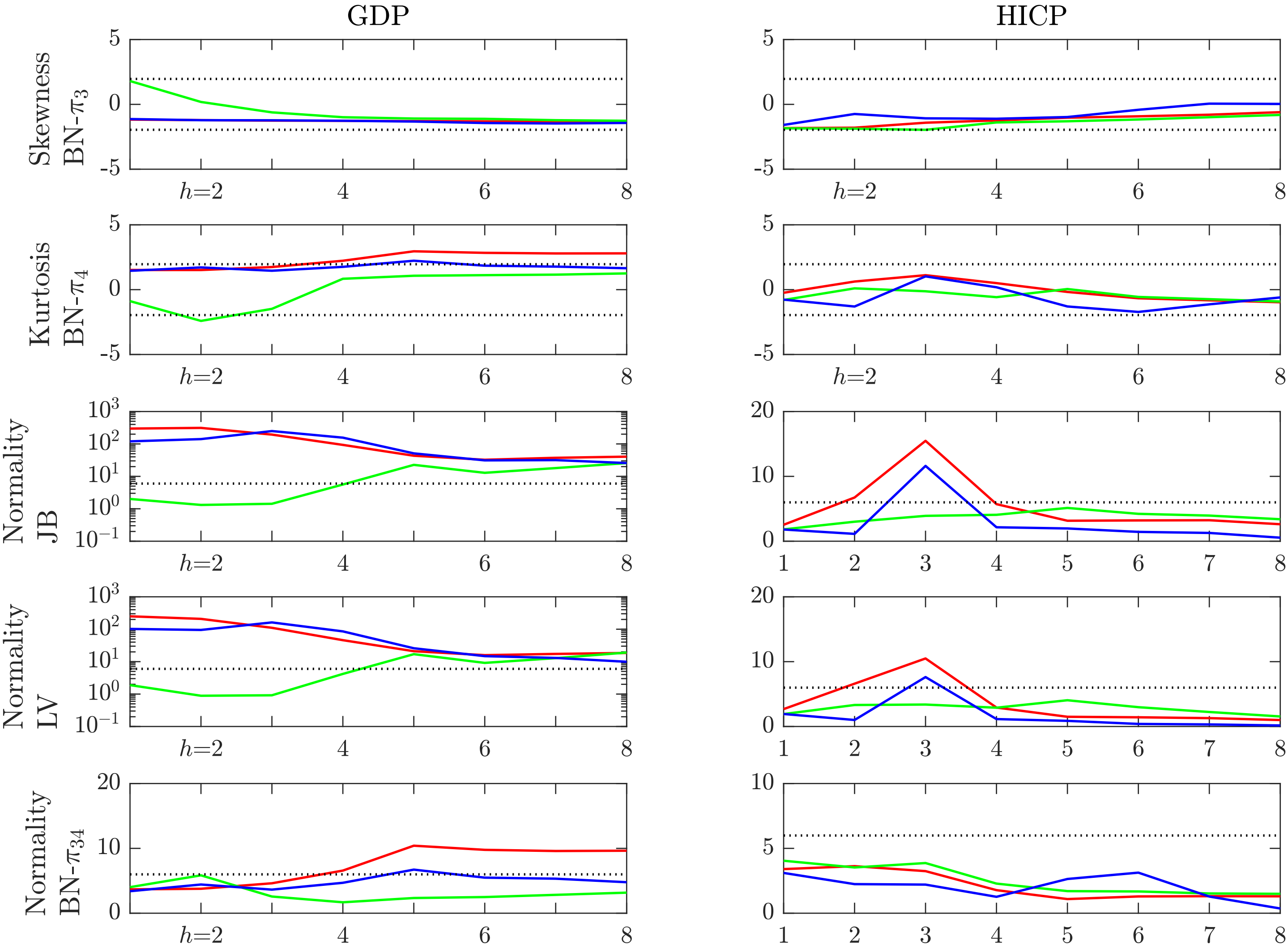

Figure 5: Normality tests

Notes: JB: Jarque and Bera (1987), LB: Lobato and Velasco (2004), BN (π_3, π_4 and π_{34}): Bai and Ng (2005) tests for skewness, kurtosis and normality. Red line: test statistics over full sample (2001Q4-2016Q3), green line: excluding financial crisis (2008Q1-2009Q2), blue line: adjusted errors test statistics over full sample. Dotted lines show critical values at 5% significance level. Y-axis is linear unless otherwise specified as logarithmic.

table as Table 8. ΔY represents the change in the underlying variable (i.e. *change* in GDP growth or *change* in inflation) and ΔF the change in the forecast, and the input in the cells is the number of times the condition is satisfied (e.g. upper-left cell when both ΔY and ΔF are positive). In this context, the success rate is defined as the number of times the realised path of the variable of interest has been correctly anticipated by the (B)MPE forecaster, as a share of the total number of forecasts. That is, the success rate is the sum of the diagonal elements of the contingency table above as a share of the total (low-right cell of the table) - as in equation 19:

$$Success\,rate(\%) = \frac{N_{11} + N_{22}}{N} \times 100 \tag{19}$$

Statistical tests evaluate whether the success rate is significantly different from 50%, i.e. different from a random, "coin toss" forecast in terms of direction. Timmerman (PT92, 1992) propose a non-parametric test for the evaluation of the directional predictive performance which, however, is derived on the explicit assumption of absence of serial correlation in the sample of the forecasts and the outcomes. In a following work, the same authors - Pesaran and Timmermann (PT09, 2009) - proposed a new test that is robust to serial correlation in the data. The properties of various tests of directional accuracy in small samples is analysed in Blaskowitz and Herwartz (2014).

In our analysis we employ PT09, the classical PT92 for reference purposes and a dynamic regression-based approach as in Blaskowitz and Herwartz (2014):

$$\tilde{y}_t = c + \beta \tilde{f}_t + \sum_{i=1}^{p} \gamma_i \tilde{y}_{t-i} + \sum_{j=1}^{p} \delta_j \tilde{f}_{t-j} + u_t \tag{20}$$

where $\tilde{y}_t$ and $\tilde{f}_t$ are dummy variables that take values of 0 or 1 to denote downward (zero) or upward (one) movements of the change in the outcome y_t and forecast f_t respectively (e.g. when $\Delta Y > 0 \therefore \tilde{y}_t = 1)$. We allow for all potential combinations of lagged values for $\tilde{y}_t$ and $\tilde{f}_t$ by not restricting $i = j$ for each regression. The optimal number of lags for $\tilde{y}_t$ and $\tilde{f}_t$ is selected using the Akaike Information Criterion (AIC) out of a maximum of four lags.

We consider tests of the null hypothesis of zero covariance between realisations and forecasts: $H_0 : cov(\Delta Y, \Delta F) = 0$. Directional accuracy holds when we reject the null. In the regression-based test, directional accuracy is tested by conducting a t-test of the null hypothesis $H_0 : \beta = 0$ using HAC standard errors (rejecting the null would imply directional accuracy in the forecasts). In particular, robust standard errors are derived with a Bartlett kernel and the lag selection parameter is set to the integer part of $4 \times (T/100)^{2/9}$ as described in Newey and West (1994).

Table 9 provides the results of this analysis[18]. The success rate ranges between 50% - 70% indicating that the realised paths of GDP and inflation were correctly anticipated more often than not. The only exception is the GDP growth forecast at $h = 8$ where the success rate is 44%. In terms of test statistics, GDP forecasts at $h = 2$ and 4 appear to be directionally accurate. With regard to the GDP $h = 8$ forecast error the results should be interpreted with care as the tests indicate directional accuracy, however, in the wrong direction. Indeed, Paloviita et al. (2017) find that the correlation of actual GDP growth and medium-to-long term GDP-forecasts of Eurosystem/ECB staff is actually negative, albeit very low. As regards inflation, the results are not very encouraging. Although success rates are higher than 50% they are not statistically significantly different from 50%, especially at $h = 4, 8$ whereas even at the very short horizon $h = 2$ the null is rejected only at 10% significance level. Interestingly, success rates of GDP y-o-y growth forecasts are higher to those of inflation at $h = 2, 4$, even though inflation data is much more timely than GDP data.

Correcting for errors in the conditioning assumptions significantly improves the directional accuracy of the Eurosystem/ECB forecasts, especially for inflation. Success rates increase and the null hypothesis of zero covariance between forecasts and realisations is rejected at conventional levels for GDP in all but one cases (GDP adjusted forecasts $h = 8$ under the dynamic regression test) and at 1% level in all cases for HICP inflation, providing robust evidence of directional accuracy of the (B)MPE forecasts once the errors in the conditioning assumptions are taken into account. The sharp improvement in the adjusted inflation forecasts should be seen against the worsening of those in the unbiasedness tests. We mentioned in Section 4.1.1 above some reasons for this outcome, for example possibly different models/elasticities used for short-term inflation projections and that the BMEs might not be a good representation of the forecasting model or even accurate estimates of the impacts of changes in assumptions to the forecasts. Same reasoning could also apply in the context of the directional accuracy tests but, since now the results improve, those would be reasons for a cautious interpretation of this improvement. As mentioned earlier in the case of bias, it would thus be interesting to allow for alternative models

[18] Note the tests are done for $h = 2$ instead of $h = 1$ to allow for a change in the projected variable.

Table 9: Directional accuracy tests

	GDP	HICP	GDP-adj.	HICP-adj.
		$h = 2$		
Success rate(%)	77.6	63.8	81.0	77.6
PT92 p-value	0.00***	0.04**	0.00***	0.00***
PT09 p-value	0.00***	0.06*	0.00***	0.00***
Dyn. regr. p-value	0.00***	0.06*	0.00***	0.00***
		$h = 4$		
Success rate(%)	66.1	57.1	73.2	75.0
PT92 p-value	0.01***	0.26	0.00***	0.00***
PT09 p-value	0.01***	0.43	0.00***	0.00***
Dyn. regr. p-value	0.07*	0.28	0.08*	0.00***
		$h = 8$		
Success rate(%)	44.2	51.9	67.3	80.8
PT92 p-value	0.04**	0.65	0.01**	0.00***
PT09 p-value	0.02**	0.61	0.03**	0.00***
Dyn. regr. p-value	0.03**	1.00	0.32	0.00***

PT92/PT09: Pesaran and Timmermann (1992) and (2009) tests of directional accuracy. The dynamic regression refers to equation 20 and we report the p-value of β. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

to correct the forecast for the errors in assumptions and compare the results.

We close this section by looking at the results of these tests in other studies in the literature. Melander et. al (2007) find that the EC forecasts are generally directionally accurate, with success rates of 80%-90% for current-year forecasts and 60%-80% of year-ahead forecasts. These numbers refer to both GDP and inflation forecasts and are for the EU countries and the EU as a whole. They are in general higher to the success rates of the (B)MPE but these are forecasts at annual frequency. The success rates are statistically significant with low p-values but tested only with the simple χ-square test that does not account for serial correlation in the data. Pain et al. (2014) find that the OECD GDP and inflation forecasts are also directionally accurate, focusing over the period 2007-2012, with success rates similar to those of the EC (Melander et al., 2007). The null hypothesis that the projections and outcomes are independent is strongly rejected at 1% significance level using the PT92 test; similarly not accounting for the inherent correlation in the forecast errors.

5.3 Performance against benchmark models

Forecast performance is also tested against two simple benchmark models: the Random Walk (RW), also called the “naive forecast” and the AR(1). This type of comparison is common in the literature - especially against the RW - but it should be interpreted with care. First, the forecasts performed by the benchmark models are not restricted by any type of assumptions - they are unconditional forecasts - in contrast to the (B)MPE which is a conditional forecast on a set of assumptions. The tests performed do not account for the potential error in the conditioning variables and the uncertainty incurred by these conditions, thereby giving some advantage to the benchmark models. Naturally, to the extent that the adjusted error accurately represents the forecast error that would be made by conditioning on the actual, ex-post, path of the conditioning variables, then this caveat is reduced[19]. In that sense, the comparison between the forecast error made by the simple models against the (B)MPE adjusted error is more valid. Second, although a unique (B)MPE projections model does not exist, it is natural to assume that it is much more complicated than the simple benchmarks. West (1996), West and McCracken (1998) and Clark and McCracken (2001) have shown that in performing forecast comparisons it is important to reflect the uncertainty in the parameter estimation error; an issue that becomes more relevant for bigger and more complicated models. In the current context, it would be impossible to correct for this uncertainty since, as mentioned, a single (B)MPE model does not exist and, as such, the RW and AR(1) models are not nested models for the (B)MPE process[20]. Nevertheless, the significantly higher degree of complexity of institutional forecasts should be seen as another factor contributing to a more favourable forecasting performance for the RW and the AR(1). Last but not least, the RW/AR(1) models are able to forecast one single variable at a time, whereas the Eurosystem/ECB forecasts incur a battery of variables forecasted simultaneously ensuring consistency of those variables and, importantly, a coherent economic story - a valuable input to the policy maker.

With these caveats in mind, we shall now move on to describe how the comparison is performed. Eurosystem/ECB versus the benchmark model is tested for equal predictive accuracy using the Diebold and Mariano (DM, 1995) test with a small-sample adjustment as in Harvey, Leybourne and Newbold (HLN97, 1997). We define a sequence as the difference between the squared forecast errors:

$$d_t^{DM} = (e_t^{(B)MPE})^2 - (e_t^{Benchmark})^2 \tag{21}$$

A one-sided (lower-tail) test of equal predictive performance of the Eurosystem/ECB against

[19] For this to be true, it should be that the BMEs represent a good proxy of the main forecasting model. Nevertheless, the Eurosystem/ECB projections are not the outcome of a single model.

[20] For an evaluation of BoE's suite of pure statistical, judgment-free models see Kapetanios et al. (2008). Groen et al. (2009) assess BoE's inflation and GDP growth forecasts versus univariate and VAR models.

the benchmark is thus:

$$\begin{aligned} H_0 &: E[d_t^{DM}] = 0 \\ H_1 &: E[d_t^{DM}] < 0 \end{aligned} \tag{22}$$

The test is formalised by regressing d_t on a constant c and testing whether this is zero under the null ($c = 0$) or negative under the alternative ($c < 0$) using HAC standard errors and a small-sample adjustment as in HLN97:

$$d_t^{DM} = c + u_t \tag{23}$$

We also perform a second type of test, based on the notion of forecast encompassing. Following Elliott and Timmermann (2016, p. 393), assuming a Mean Square Error loss function and two competing forecasts f_1 and f_2 under the information set Ω_t, then f_1 is said to encompass f_2 when:

$$E_y[(y - f_1))^2 | \Omega_t] \leq \min_w E_y[(y - ((1 - w)f_1 + wf_2))^2 | \Omega_t] \tag{24}$$

for some weight $w \in \mathbb{R}$. Based on the above, Harvey, Leybourne and Newbold (HLN98, 1998) suggest the following sequence to be tested whether it is significantly different from zero, in a similar way as in (23):

$$d_t^{HLN98} = (e_t^{(B)MPE}) \times (e_t^{(B)MPE} - e_t^{Benchmark}) \tag{25}$$

However, HLN98 test explicitly assumes zero-mean errors and thus, as suggested in Elliott and Timmermann (2016, p. 395), we employ Marcellino's (2000) procedure to adjust the test statistic. In more detail, we substitute the errors with the estimated residuals $\hat{u}_t$ of the MZ regression in the unbiasedness section (see equation 7) - for both the (B)MPE and the benchmark-model forecast - and transform (25) into:

$$
\begin{aligned}
&y_t^j = c_j + \beta_j f_t^j + u_t^j \text{ }, j = \{(B)MPE, Benchmark\} \\
&\tilde{d}_t^{HLN98} = (\hat{u}_t^{(B)MPE}) \times (\hat{u}_t^{(B)MPE} - \hat{u}_t^{Benchmark})
\end{aligned}
\tag{26}
$$

The encompass one-sided (upper-tail)[21] hypothesis test is thus:

$$
\begin{aligned}
&H_0 : E[\tilde{d}_t^{HLN98}] = 0 \\
&H_1 : E[\tilde{d}_t^{HLN98}] > 0
\end{aligned}
\tag{27}
$$

and is performed in a similar manner as in (23). In other words, encompassing, under the null, requires a zero correlation between the (B)MPE forecast errors and the difference of the (B)MPE and the benchmark forecast errors. If the null of zero correlation is rejected, then the benchmark model is informative for the (B)MPE forecast.

Overall, two tests are performed as outlined in equations (23) and (27) and the results are provided in Table 10 for the RW and Table 11 for the AR(1) benchmark, together with the ratios of the RMSEs of the (B)MPE against the benchmark models[22]. The ratios of the RMSEs are below unity and the constant of the regression in (23) is negative, indicating in general better forecasting accuracy of the (B)MPE (for GDP forecast at $h = 8$ against the AR(1) the RMSE ratio is 1). The RMSE ratio decreases when computed against the adjusted (B)MPE error, and in some cases substantially so, indicating that errors in assumptions weigh heavily on forecasting performance. In terms of statistical significance, GDP growth and inflation forecasts appear to be significantly better than the simple benchmarks at conventional significance levels, especially at short horizons. At $h = 8$, however, the superiority of the (B)MPE forecasts against the AR(1) cannot be confidently established as the results of the two tests disagree. The DM test provides no evidence of significantly superior performance of the (B)MPE but, at the same time, the forecast encompassing test does not provide evidence of the AR(1) being informative for the (B)MPE forecast.

[21] One sided hypothesis as proposed in HLN98 holds when negative weight w in (24) is not allowed, i.e. $0 \leq w \leq 1$. This should not hold necessarily under MSE loss - see Elliott and Timmermann (2016, p. 312-314)

[22] The models are estimated in quarter-on-quarter growth rates and we use the y-o-y growth rate forecast.

Table 10: Eurosystem/ECB relative forecasting performance to RW

	GDP	HICP	GDP-adj.	HICP-adj.
	$h = 1$			
RMSE ratio	0.80	0.23	0.65	0.20
DM constant	-0.13	-0.11	-0.21	-0.11
DM p-value	0.06*	0.00***	0.01***	0.00***
Encompass constant	-0.03	0.00	-0.06	0.00
Encompass p-value	0.95	0.91	0.97	0.54
	$h = 4$			
RMSE ratio	0.67	0.64	0.48	0.36
DM constant	-3.04	-0.96	-4.25	-1.42
DM p-value	0.04**	0.05*	0.05**	0.03**
Encompass constant	-0.09	0.00	-0.13	0.00
Encompass p-value	0.72	0.50	0.69	0.51
	$h = 8$			
RMSE ratio	0.68	0.66	0.48	0.50
DM constant	-5.33	-1.41	-7.77	-1.87
DM p-value	0.09*	0.09*	0.07*	0.11
Encompass constant	0.19	0.00	0.02	-0.01
Encompass p-value	0.12	0.52	0.48	0.52

DM constant refers to eq. 23 and encompass constant to eq. 27. DM and encompass p-values are estimated using HLN97 small sample size adjustment. HAC (Bartlett) standard errors with bandwidth set to $h - 1$. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

Table 11: Eurosystem/ECB relative forecasting performance to AR(1)

	GDP	HICP	GDP-adj.	HICP-adj.
	$h = 1$			
RMSE ratio	0.76	0.27	0.62	0.23
DM constant	-0.19	-0.08	-0.28	-0.08
DM p-value	0.01***	0.00***	0.01**	0.00***
Encompass constant	-0.07	0.00	-0.09	0.00
Encompass p-value	0.98	0.86	0.98	0.61
	$h = 4$			
RMSE ratio	0.79	0.79	0.57	0.44
DM constant	-1.63	-0.41	-2.93	-0.87
DM p-value	0.08*	0.08*	0.09*	0.01**
Encompass constant	-0.14	0.02	-0.12	0.02
Encompass p-value	0.79	0.36	0.67	0.44
	$h = 8$			
RMSE ratio	1.01	0.83	0.71	0.63
DM constant	0.14	-0.51	-2.50	-0.96
DM p-value	0.72	0.22	0.10	0.17
Encompass constant	0.22	0.11	-0.07	-0.06
Encompass p-value	0.16	0.20	0.56	0.62

DM constant refers to eq. 23 and encompass constant to eq. 27. DM and encompass p-values are estimated using HLN97 small sample size adjustment. HAC (Bartlett) standard errors with bandwidth set to $h - 1$. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

5.3.1 Time-varying relative forecasting performance

So far the performance of the (B)MPE against the benchmark models has been evaluated over the entire sample and the conclusion of overall superior forecasting performance of the (B)MPE applies to the whole sample, on average. It is interesting however to see how this performance has evolved over time. It can be of particular interest to know whether there have been certain points in time, or over the cycle, over which the benchmark model(s) have outperformed the (B)MPE.

We test the forecasting performance of the (B)MPE against the benchmark models over time by performing the Giacomini and Rossi (GR, 2010) fluctuation test. In a nutshell, we compare the relative forecasting performance of the two competing models by checking whether the loss differential - i.e. the difference in the squared forecast error - over rolling windows is significantly different from zero at any point in time. At forecast horizon h, for a given total sample size of forecast errors P_h, we set the rolling-window size m such that $m/P_h = 0.3$. The critical values are then obtained from Giacomini and Rossi (2010), Table 1. With this approach, for a one-sided test at 5% the critical value is set to -2.770 at all horizons. The results are reported in Figure 6[23]. The loss differential is defined in a similar manner as in equation (21) such that a negative value indicates a better forecasting performance of the (B)MPE on average over the period in question.

Overall, the results are somewhat less supportive of a superior forecasting performance of the Eurosystem/ECB forecasts against the benchmark models that was reported for the overall sample. For GDP at $h = 1$ the fluctuation test shows that the (B)MPE significantly outperformed the benchmark models only for very short time intervals in the period following the sovereign crisis. At longer horizons, the (B)MPE forecast is in general not significantly better than the benchmark models - with some exceptions at $h = 8$ against the RW. On a more positive note, though, the loss differential is most often negative implying in general better forecasting performance of the (B)MPE, albeit not always significantly so. The loss differential is however often positive against the AR(1) at $h = 8$, implying better forecasting performance of the benchmark model (although not tested for statistical significance). For inflation, the (B)MPE outperformed the benchmark models at $h = 1$ on several occasions except during the financial and sovereign-debt crises and the inter-crises period. Interestingly, this also includes the recent

[23] The analysis starts later due to lack of data in the early parts of the sample, necessary to estimate the benchmark models.

period of low inflation over which the Eurosystem/ECB staff have persistently over-predicted inflation. Nevertheless, the results are less favourable for the (B)MPE forecaster at longer forecasting horizons of inflation. The (B)MPE significantly outperformed the benchmark models only at few, sporadic cases although against the RW the loss differential is generally negative. Before and during the financial crisis it appears that the AR(1) model was performing better than the (B)MPE at $h = 4, 8$. This is the period in which a significant and persistent tendency to under-predict inflation was reported in the time-varying unbiasedness tests of Section 4.1.3. In line with the decrease in the time-varying bias, Figure 6 shows clearly an improvement of the inflation forecasting performance of the Eurosystem/ECB staff against the benchmark models through time (for $h = 4, 8$ except RW/ $h = 4$). Especially against the AR(1), the loss differential depicts a clear downward trend.

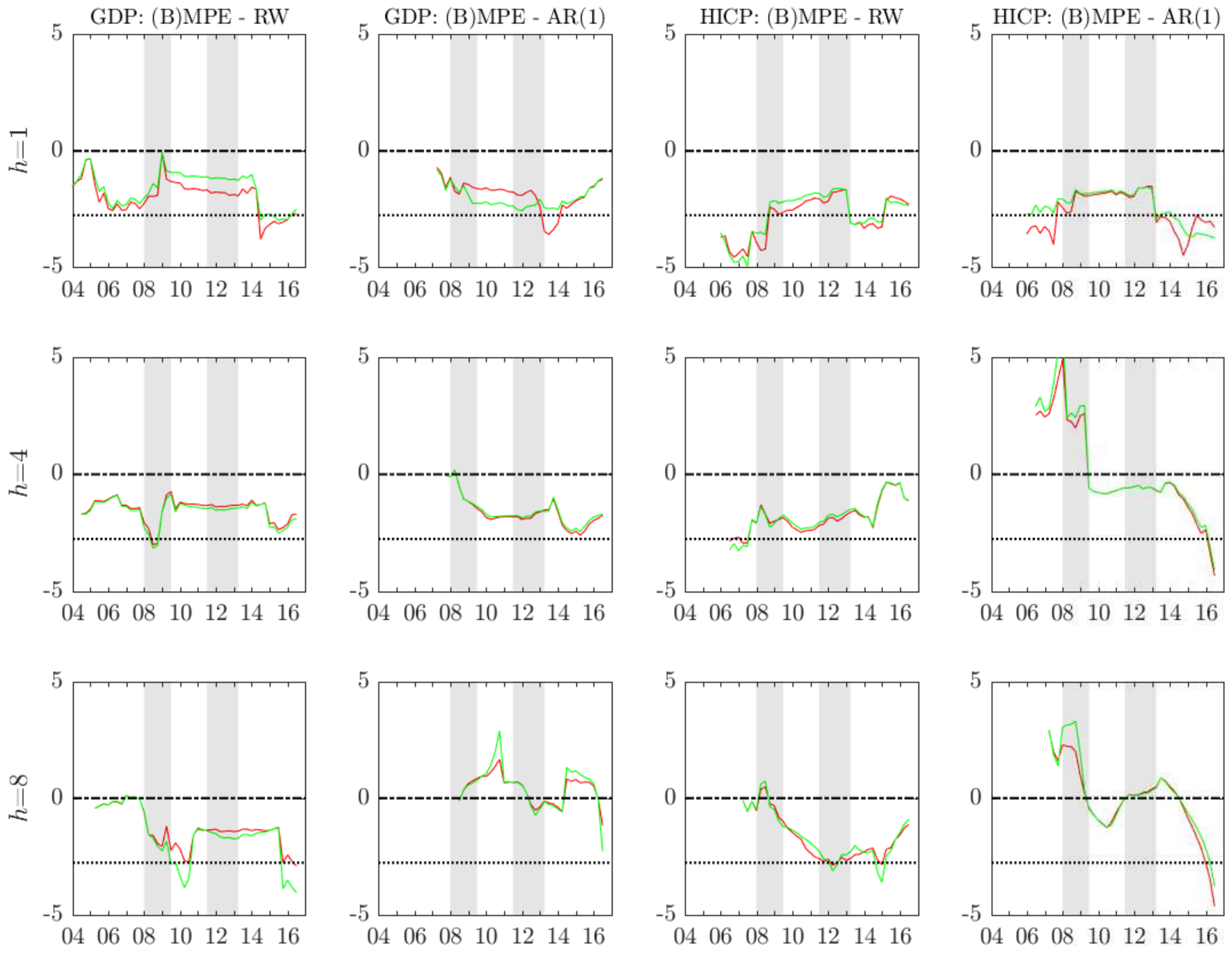

Figure 6: Time-varying Eurosystem/ECB relative forecasting performance to benchmark models

Notes: Giacomini and Rossi (2010) one-sided fluctuation test. Green line: bandwidth $h-1$, red line: bandwidth set according to Newey and West (1994). Dotted-lines: critical values at 5% significance level.

6 Forecasting performance against other forecasters

In this section we show how the Eurosystem/ECB projections compare against the forecasts of other institutions - European Commission, IMF, and OECD, as well as the private sector - Survey of Professional Forecasters (SPF). The sample includes calendar-year forecasts done over 2000-2016.

6.1 Performance of calendar-year forecasts

It is not feasible to make a direct comparison of all the above mentioned forecasters' performance in calendar-year forecasts due to a series of reasons. First of all, the sample in calendar-years is rather short since Eurosystem/ECB existence and thus it is not safe to conclude on any robust results and significance. Additionally, the forecasts are not published simultaneously by all the institutions and thus these differences in timing represent different information sets. This, ceteris paribus, would result in a variation of forecasting performance. Furthermore, slightly different definitions of the euro area in terms of the aggregation methods used could also explain differences in the forecasts. Finally, similarly to the (B)MPEs other institutions presumably condition their projections on a set of exogenous variables which could be different than the Eurosystem/ECB.

With these caveats in mind, we compare BMPE performance vis-a-vis each forecaster following the approach in ECB (2013). That is, we take the average of the current-year and next calendar-year forecasts of the projections produced in the 2^{nd} and the 4^{th} quarter of each year. Thus, in the case of the Eurosystem only the BMPEs are used. Other forecasters produce forecasts within the second quarter of each year. For example SPF's and IMF's projections are published in April, OECD's in May, whereas BMPEs in June. The RMSEs of the current-year and next calendar-year forecasts are presented in Figure 7. Table 12 shows the RMSE ratios as well the Diebold-Mariano values following Harvey et al. (1997). Despite the rather small sample, HLN97 show that for total sample size $N = 16$ the adjusted DM test is relatively correctly sized.

The RMSEs bars show that the forecasting performance of all forecasters is very similar and that the Eurosystem forecasts compare very favourably with the other forecasters. To some extent, though, this is to be expected given the more favourable timing of the BMPE forecasts (e.g. the forecast done in December when a lot of information about the year is available). In particular, current-year and next calendar-year GDP forecasts feature lower RMSEs than those of other forecasters, more clearly so against the European Commission (EC) and the IMF. Inflation

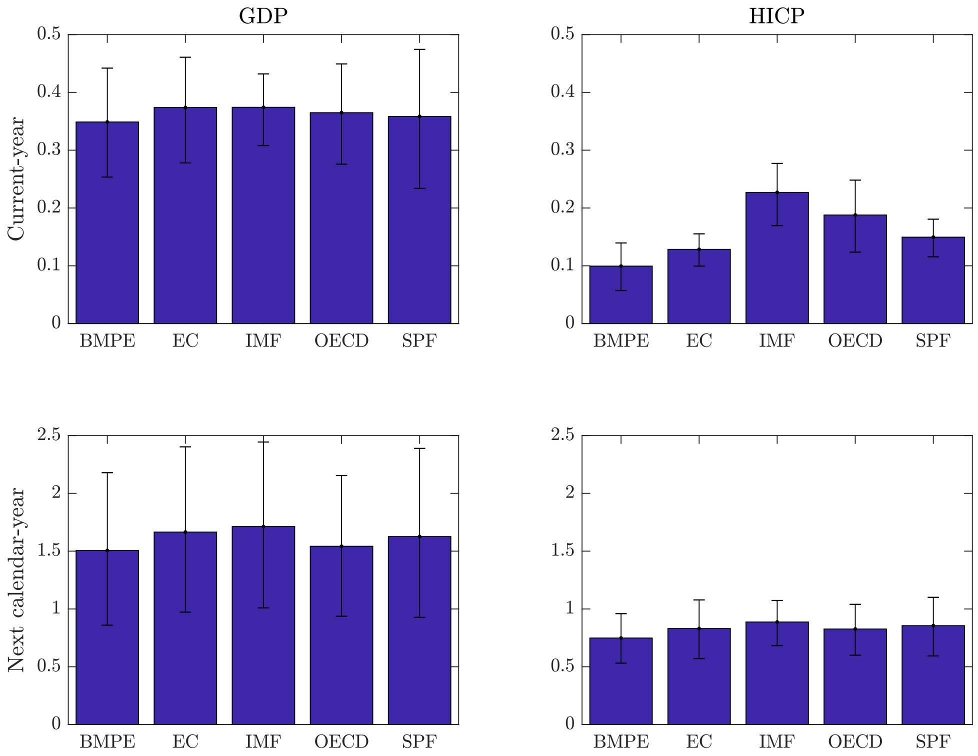

Figure 7: RMSEs: Eurosystem/ECB and other forecasters
Notes: Current-year and next calendar-year RMSEs estimated as the average of projections produced in the second and the fourth quarter of each year. 95% confidence intervals are estimated using percentile bootstrap employing the stationary block-bootstrap of Politis and Romano (1994) with block size set as in Politis and White (2004) and with 9999 repetitions.

forecasts of the Eurosystem staff are even more clearly superior to other forecasters by a simple visual inspection of the relevant figures. Similar conclusions regarding inflation were obtained in a similar analysis in ECB (2013), whereas GDP forecasts appear to have improved somewhat over the last four years especially against the IMF forecasts.

The RMSE ratios are always inferior to unity, indicating overall a better forecasting performance of the Eurosystem, although in terms of statistical significance the conclusions are somewhat mixed (Table 12). For the current-year forecasts, no conclusions can be made about significantly better forecasting ability of euro area GDP by the Eurosystem staff; but this can be confidently concluded for inflation and indeed against all forecasters. For the next calendar-year, the GDP forecast is significantly better against all forecasters except the OECD - at least at 10% significance level - whereas for HICP inflation only against the IMF and the SPF.

Table 12: Eurosystem relative forecasting performance to other forecasters

	EC	IMF	OECD	SPF
	GDP - Current-year			
RMSE ratio	0.93	0.93	0.96	0.97
DM constant	-0.02	-0.02	-0.01	-0.01
DM p-value	0.38	0.47	0.27	0.55
	HICP - Current-year			
RMSE ratio	0.77	0.44	0.53	0.66
DM constant	-0.01	-0.04	-0.03	-0.01
DM p-value	0.05*	0.00***	0.09*	0.01***
	GDP - Next calendar-year			
RMSE ratio	0.90	0.88	0.98	0.93
DM constant	-0.51	-0.67	-0.11	-0.38
DM p-value	0.07*	0.03**	0.43	0.09*
	HICP - Next calendar-year			
RMSE ratio	0.90	0.84	0.91	0.87
DM constant	-0.13	-0.23	-0.12	-0.17
DM p-value	0.18	0.01**	0.38	0.09*

DM constant refers to eq. 23. DM p-values are estimated using HLN97 small sample size adjustment. HAC (Bartlett) standard errors with bandwidth set to $h-1$. DM test is two-sided. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

6.2 Eurosystem/ECB relative forecasting performance to SPF

In this section we compare Euroystem/ECB against the SPF in quarterly and monthly frequency for GDP and HICP respectively. Eurosystem produces monthly frequency forecaststs for HICP and all its sub-components via the NIPE for a horizon up to 12 months ahead which are subsequently aggregated into quarterly frequency and are included in the (B)MPEs (see Section 2).

It is important to have in mind the respective cut-off dates for Euroystem/ECB and SPF. GDP forecasts produced in the (B)MPE are published at the beginning of the third month (m3) of quarter t and are produced during the second month of the same quarter (m2). Thus, the last observed data point is that of quarter $t-1$[24]. Similarly, NIPE is conducted in the second month of quarter t when the last observed value is that of the first month of the same quarter (m1). The SPF survey is conducted by the ECB in the first month of each quarter and includes the 4- and 8-quarters ahead forecasts for GDP and the 12- and 24- month ahead forecasts for HICP. When the private forecasters are asked to provide the previously mentioned forecasts the last quarterly GDP figures that are available are that of quarter $t-2$ and the last monthly HICP

[24] This is the case since the introduction of the preliminary flash by the Eurostat on 29 April 2016, published one month after the end of the quarter. The flash release, though, arrives 45 days after the end of the quarter which would make it rather close to the cut-off date. We leave for further research to decompose forecasters performance to individuals' ability and the effect of informational advantage according to Andersson et al. (2017).

figure is - in the best case - that of the last month of the quarter preceding the survey.

Therefore, in order to be as fair as possible with respect to the information set available to the two set of forecasts regarding the GDP we compare SPF's forecasts produced in quarter t with (B)MPE forecasts produced in quarter $t-1$. For HICP forecasts, however, is not feasible to make a direct comparison of the same projection horizon and therefore the NIPE 11-month ahead forecasts are compared with the SPF 12-month ahead forecasts. The following table shows clearly the timing of the forecasts.

Table 13: (B)MPE/NIPE and SPF cut-off dates

Quarter Q, Months **m**		
m1	**m2**	**m3**
SPF	(B)MPE/NIPE	-
$\text{GDP}_q(q-2)$	$\text{GDP}_q(q-1)$	-
$\text{HICP}_m(m-1)$	$\text{HICP}_m(m-1)$	-

We first perform a full sample analysis from 1999Q4 - 2016Q3, which corresponds to 1999Q1 (B)MPE/NIPE and the 1999Q2 SPF survey. Table 14 provides the results of the HP tests for unbiasedness - for the sake of brevity we only focus on this test. Similarly to Section 4.1 above, we find evidence of bias in the GDP forecasts of the ECB/Eurosystem staff, especially at long horizons. Based on this test, the bias of the SPF respondents is very similar, both in terms of direction, magnitude and statistical significance. That is, the bias at 4 and 8 quarters ahead horizon is negative (over-prediction), they are of almost the same size - approximately -0.4% at $h = 4$ and -1% at $h = 8$, borderline (non-) significant at $h = 4$ and statistically very significant at $h = 8$. For inflation, both forecasters tend to under-predict - again by an impressive similarity - without signs of a clear persistent and statistically significant bias.

In terms of forecasting accuracy, the performance of the two forecasters is rather similar and there are only mixed signs of superior performance. RMSE ratios, DM and encompassing tests are provided in Table 15. Due to the caveats above the DM test is two-sided. The RMSE ratio is close to unity, slightly more accurate forecasts are given by the SPF respondents in the case of GDP at $h = 4$ and somewhat worse for inflation. For the case of inflation, the DM test does not provide enough evidence that the forecasts are statistically different, while at the same time the encompassing test does not suggest that the (B)MPE could be informed by the SPF. For GDP at long horizons, although the RMSE ratio is very close to unity the DM test suggests that the

Table 14: Unbiasedness - HP test: Eurosystem/ECB and SPF

	(B)MPE GDP	SPF GDP	NIPE HICP	SPF HICP
	$h = 4$ quarters		$h = 12$ months	
Bias	-0.39	-0.35	0.16	0.15
P-value (bw $h - 1$)	0.21	0.20	0.45	0.50
P-value (bw Andrews)	0.08*	0.10	0.46	0.50
	$h = 8$ quarters			
Bias	-1.09	-0.98	-	-
P-value (bw $h - 1$)	0.01**	0.02**	-	-
P-value (bw Andrews)	0.00***	0.00***	-	-

Bias refers to the value of the constant in equation 8. The p-values are calculated using HAC (Bartlett) standard errors with bandwidth (bw) set to to $h - 1$ and according to Andrews (1991). *, **, *** indicate the null hypothesis rejected at 10%, 5% and 1% significance level respectively.

two forecasts are statistically different, but again the encompassing test does not suggest that the (B)MPE can be informed from the SPF.

Table 15: Eurosystem/ECB relative forecasting performance to SPF

	GDP	HICP	GDP
	h=4 qt.	h=12 m.	h=8 qt.
RMSE ratio	1.15	0.94	1.03
DM constant	0.60	-0.10	0.29
DM p-value	0.12	0.28	0.00***
Encompass constant	0.42	0.01	0.00
Encompass p-value	0.00***	0.45	0.54

DM constant refers to eq. 23 and encompass constant to eq. 27. DM and encompass p-values are estimated using HLN97 small sample size adjustment. HAC (Bartlett) standard errors with bandwidth set to $h - 1$. DM test is two-sided. *, **, *** indicate rejection of the null at 10%, 5% and 1% significance level respectively.

In order to understand the changing dynamics over time we perform the two-sided GR test. The graphical representation of the test is provided in Figure 8. For GDP, the SPF forecasts have always been more accurate than the (B)MPE. This does not depend neither on the forecasting horizon nor on the choice of bandwidth. Nevertheless, their difference is not statistically significant at any point in time for $h = 4$ forecasts but it has been on several horizons for 8-quarters ahead forecasts: pre-2006 and since 2011-12 depending on the choice of bandwidth. On the other hand, the results for inflation are interesting: they show a continuous improvement in the forecasting performance of the (B)MPE against the SPF from the beginning of the sample to the end of it. The SPF forecasters outperformed the (B)MPE before 2008, but not significantly so. Since the beginning of the crisis the (B)MPE has been outperforming the SPF and the loss differential has remained below zero since; although statistical significance depends a lot on the choice of bandwidth. This result is in line with a continued decrease in the HICP inflation loss differential of the (B)MPE against benchmark models, as discussed in Section 5.3.1, and the declining trend in the bias in Section 4.1.3.

7 Conclusion

The Eurosystem/ECB (Broad) Macroeconomic Projections Exercises constitute a source of publicly available economic forecasts and an important input to the ECB's monetary policy. This work marks a thorough analysis of the Eurosystem/ECB projection errors using techniques widely employed in the applied literature of forecast evaluation. This adds to the list of evaluation exercises of institutional forecasts like the IMF, OECD, EC and BoE.

Overall, the results are rather encouraging for the Eurosystem/ECB forecaster but raise some concerns especially with regards to GDP forecasts at long horizons. Following standard tests and criteria, we can confidently conclude that inflation forecasts are optimal and rational. They are unbiased, weakly and strongly efficient and uncertainty in the forecasts increases with the projection horizon. Time-varying tests also suggest that inflation forecasts have been in general improving over time: the bias has been falling - although it has reached negative territory in recent years - and the relative forecasting accuracy against benchmark models - like the AR(1) - and the SPF forecast has been increasing. Albeit not being an optimal property, normality of the HICP forecast errors is also supported by our analysis. However, although the path of inflation has been in most cases correctly anticipated, we do not find enough statistical evidence that the inflation forecasts are directionally accurate.

GDP forecasts are generally optimal, but we document significant room for improvement particularly at long-term forecasting horizons. GDP forecasts up to one year are unbiased, but we cannot find enough supporting evidence of absence of bias at long-horizons where we document a strong tendency to over-predict. They are weakly efficient - except in one case ($h = 1$) - and forecasting accuracy does not decrease with the forecast horizon. Efficiency tests, nevertheless, do not suggest the full use of the information available to the forecaster. This, together with the failure of unbiasedness at long-horizons, suggest departures from rationality for GDP forecasts. On a more positive note, GDP forecasts feature well against those of other institutions and against benchmark models and have generally been directionally accurate up to one-year forecasting horizon, also in a sense of statistical significance. Yet, in the long-term the

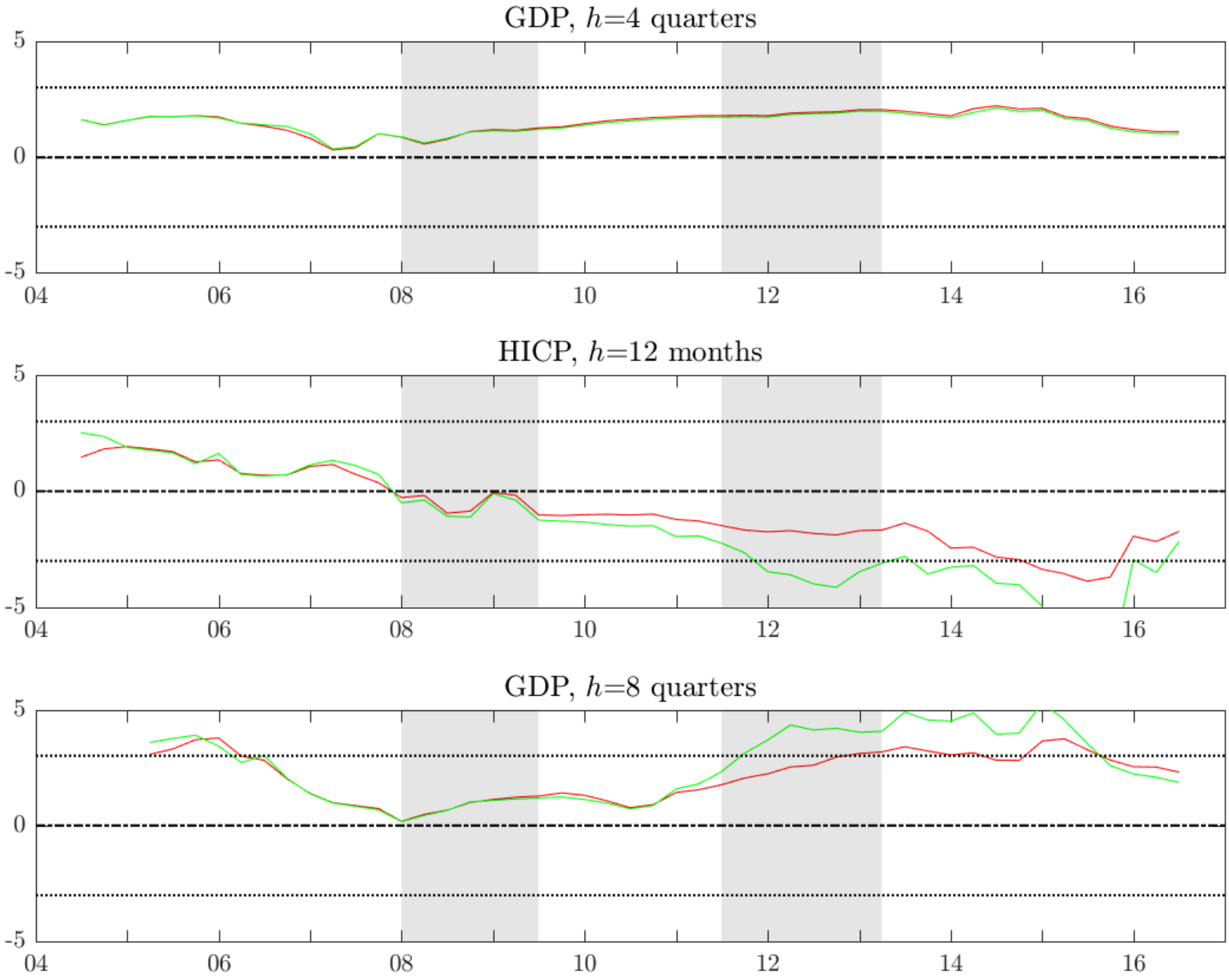

Figure 8: Time-varying Eurosystem/ECB relative forecasting performance to SPF
Notes: Giacomini and Rossi (2010) two-sided fluctuation test. Green line: bandwidth set to $h-1$, red line: bandwidth set according to Newey and West (1994). Dotted-lines: critical values at 5% significance level.

projected path was in most cases inaccurate. Further, SPF GDP forecasts tend to outperform the (B)MPE. Normality of the GDP errors also generally fails, while this can be at least partly traced to large forecast errors during the crisis.

This work opens several avenues for further research. As time goes by and data is gathered, subsequent similar analyses could further scrutinise the results. Moreover, this analysis has concentrated on euro area GDP growth and HICP inflation projections. Extending it to more variables and/or countries would provide a more comprehensive evaluation of the Eurosystem/ECB staff projections. Further, the tests are performed under the usual assumption of a mean-squared-loss function. This assumption could be relaxed and/or tested. Finally, the occasional deterioration in forecasting performance - especially the bias - after having adjusted for the errors in the conditioning assumptions is somewhat difficult to explain. Further analysis is warranted to extract the underlying reasons behind this à priori puzzling result. Since this adjustment needs to be done with a model, and is by definition model-dependent, an obvious check would be to use alternative models.

8 Appendix

8.1 Data and projection exercises

In this section we show the real-time dataset structure used for the analysis.

Table 16: Real time dataset and (B)MPEs

Dates	$2001Q4_{EA11}$	...	$t-1_{EAi}$	t_{EAi}	$t+1_{EAi}$	...	$2016Q3_{EA19}$
1990Q1	$y_{1990Q1}^{2001Q4,EA11}$	...	$y_{1990Q1}^{t-1,EAi}$	$y_{1990Q1}^{t,EAi}$	$y_{1990Q1}^{t+1,EAi}$	...	$y_{1990Q1}^{2016Q3,EA19}$
⋮	⋮	...	⋮	⋮	⋮	...	⋮
t-1		...	$f_{t-1,1}^{EAi}$	$y_{t-1}^{t,EAi}$	$y_{t-1}^{t+1,EAi}$	...	$y_{t-1}^{2001Q4,EA19}$
t		...	$f_{t,2}^{EAi}$	$f_{t,1}^{EAi}$	$y_{t}^{t+1,EAi}$	...	$y_{t}^{2001Q4,EA19}$
t+1		...	$f_{t+1,3}^{EAi}$	$f_{t+1,2}^{EAi}$	$f_{t+1,1}^{EAi}$	...	$y_{t+1}^{2001Q4,EA19}$
⋮		...	⋮	⋮	⋮	...	⋮
2016Q3		...				...	$f_{2016Q3,1}^{EA19}$

The dates in quarterly frequency in the first row of the table above correspond to the dates when each (B)MPE was conducted. Next to the date, there is an index referring to the euro area composition - i.e. the number of countries that formed the euro area - at that point in time. For example, for the column corresponding to the (B)MPE that took place in time t when the euro area composition was EAi, up to row $t-1$ there are the vintage data y as they were available for that (B)MPE. From time t and beneath for the same (B)MPE there follow the forecasts f which in the generic case vary from 9 to 12 quarters since the projection horizon for each (B)MPE ends at the second calendar year after the year when the (B)MPE was conducted (current plus 2 years). The forecast errors used in our analysis as already described in Section 3 are those which correspond to all the (B)MPEs between 2001Q4 and 2016Q3, i.e. the first and the last columns in the above table. The AR(1) forecasts compared versus the (B)MPEs in Section 3.7 have been estimated recursively with the vintage time series that were available for each (B)MPE. The estimation sample for the recursive estimation starts always in 1990Q1 as shown in the above table unless there are data availability issues.

8.2 Further figures

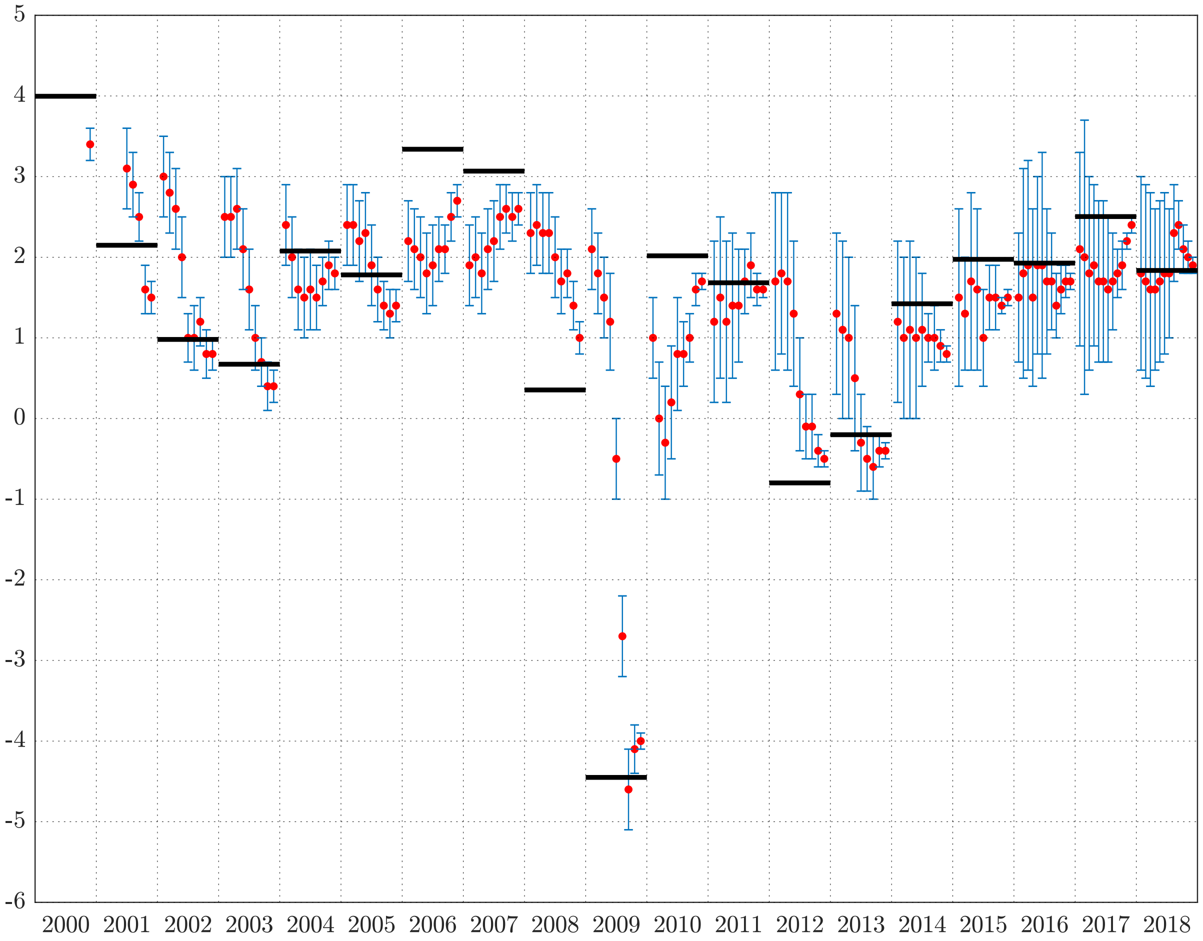

Figure 9: GDP growth published calendar-year projection ranges

Notes: Solid horizontal lines show calendar-year outcomes of the latest vintage. For each calendar-year, projection ranges (vertical bands) show how the projections for the given calendar year have evolved over consecutive (B)MPEs. The last projection range for each calendar year is produced in the December BMPE of the same calendar year. Red dots correspond to the mid-point of the projection range.

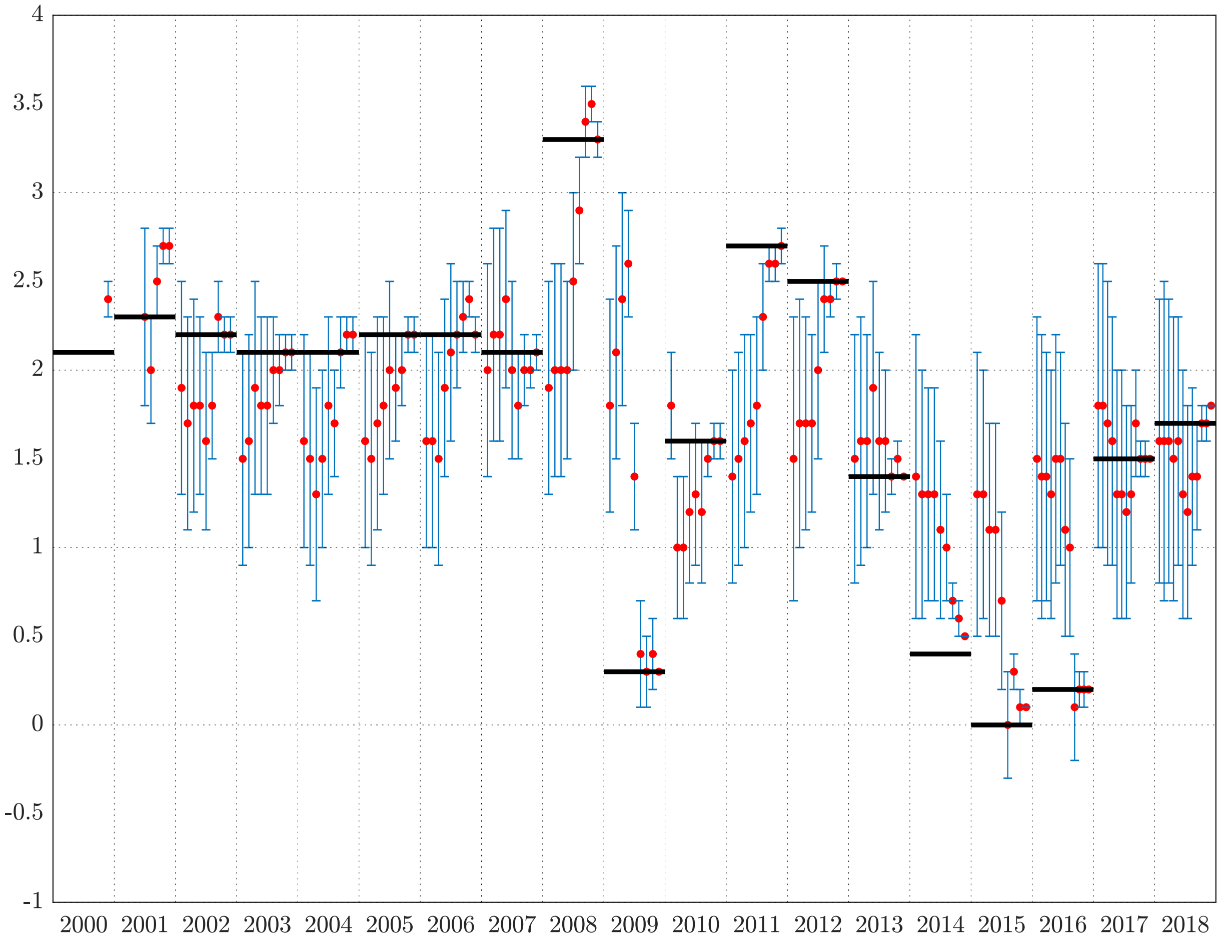

Figure 10: HICP inflation published calendar-year projection ranges

Notes: Solid horizontal lines show calendar-year outcomes of the latest vintage. For each calendar year, projection ranges (vertical bands) show how the projections for the given calendar year have evolved over consecutive (B)MPEs. The last projection range for each calendar year is produced in the December BMPE of the same calendar year. Red dots correspond to the mid-point of the projection range.

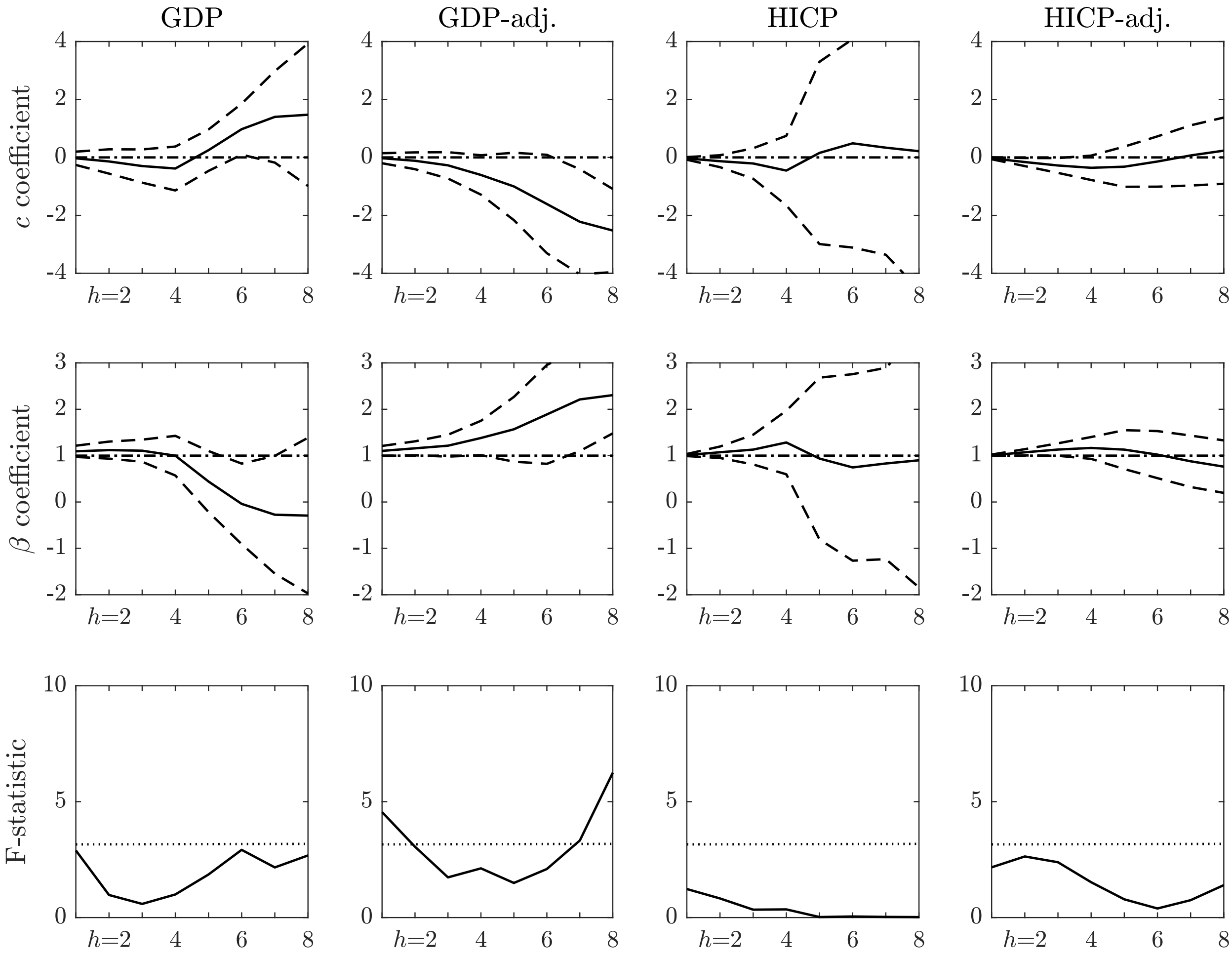

Figure 11: Unbiasedness - MZ test

Notes: First and second panel: c and β estimates of MZ equation 7 over all horizons and their 95 % confidence intervals. Bandwidth set according to Andrews (1991). Third panel: F-statistics and the critical value at 5% significance level (dotted lines) over all horizons.

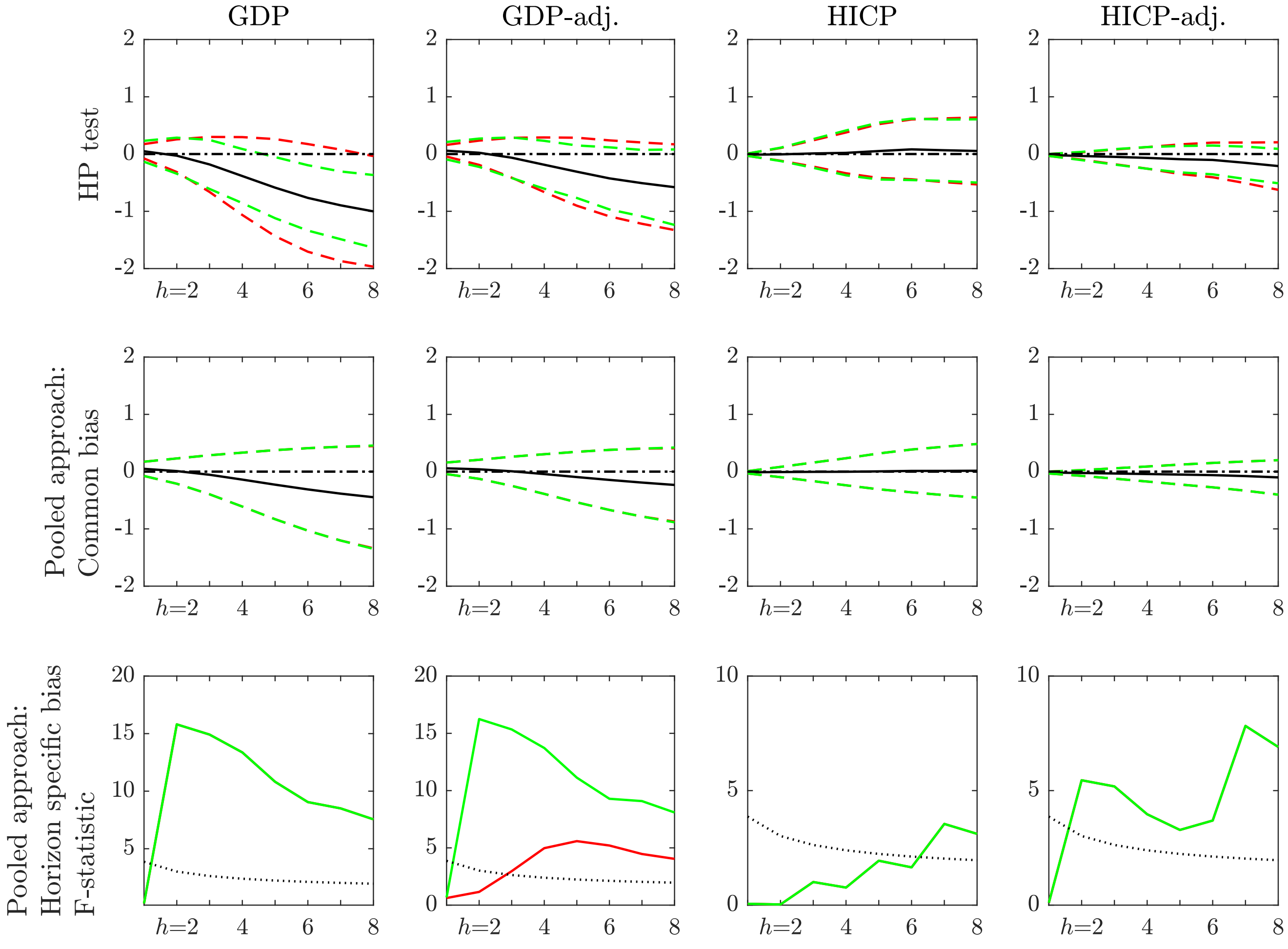

Figure 12: Unbiasedness - HP and pooled approach tests

Notes: First panel: Horizon specific HP test (equation 8) and 95 % confidence intervals (red: bandwidth set to $h-1$, green: bandwidth set according to Andrews 1991). Second panel: Clements et al. (2007) common bias test (equation 9 and 95 % confidence intervals (red: with idiosyncratic, green: without idiosyncratic shocks). Test is conducted sequentially for a common bias up to horizon h. Third panel: Ager et al. (2009) F-statistic for the joint hypothesis of zero horizon-specific bias (red: with idiosyncratic, green: without idiosyncratic shocks). Test is conducted sequentially for a horizon-specific bias up to horizon h. Dashed lines: F-test critical value at 5% significance level.

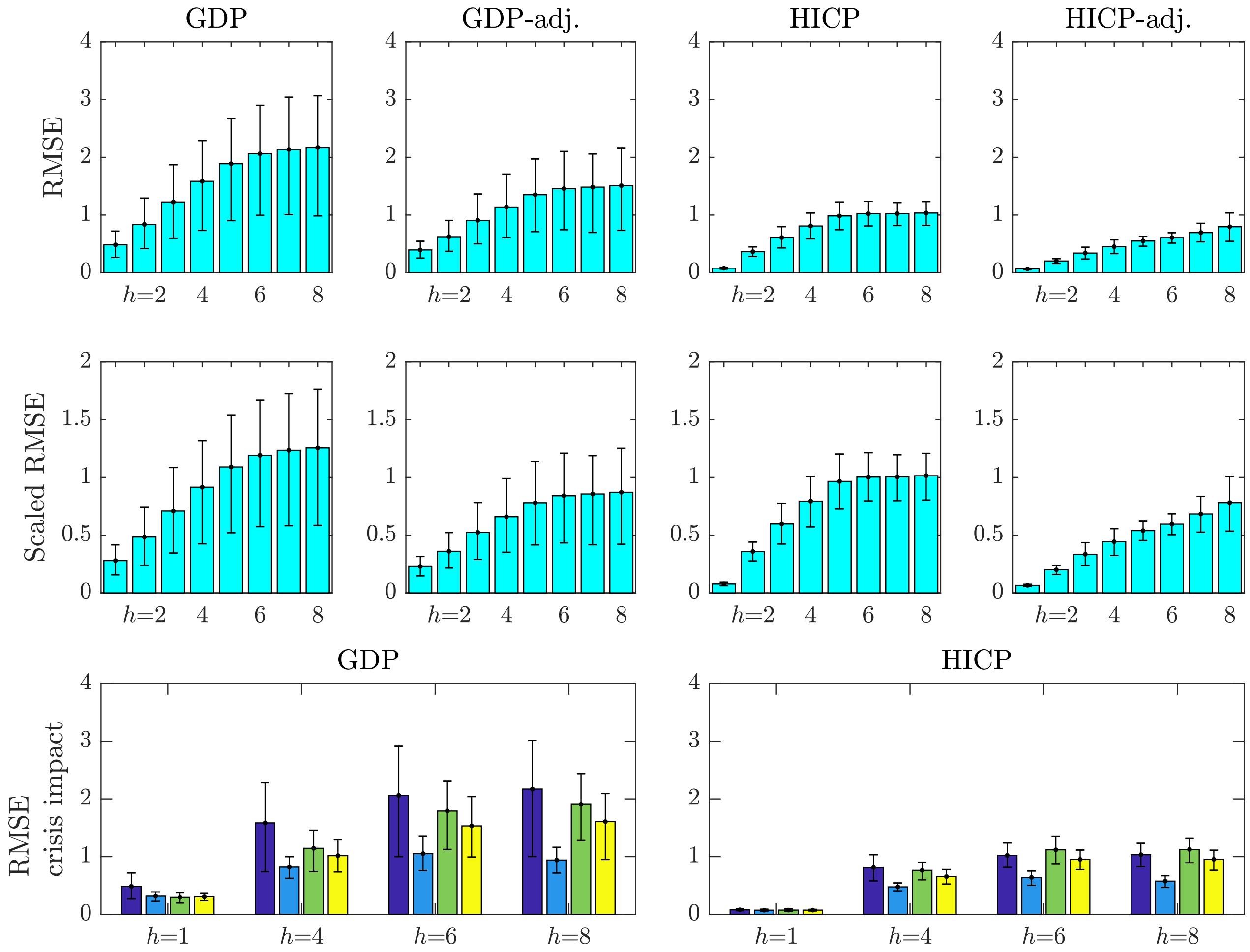

Figure 13: RMSEs

Notes: First and second panel: RMSEs and scaled RMSEs over all horizons. Third panel - RMSEs: purple: full sample (2001Q4-2016Q3), blue: pre-financial crisis (2001Q4-2007Q4), green: post-financial crisis (2009Q3-2016Q3), yellow: full sample excluding financial crisis (2008Q1-2009Q2). 95% confidence intervals are estimated using percentile bootstrap employing stationary bootstrap of Politis and Romano (1994) with block size set as in Politis and White (2004) with 9999 repetitions.

References

[1] Ager, P., Kappler, M., and Osterloh, S. (2009) *The accuracy and efficiency of the Consensus Forecasts: A further application and extension of the pooled approach.* **International Journal of Forecasting**, 25(1), 167-181.

[2] Alessi, L., Ghysels, E., Onorante, L., Peach, R., and Potter, S. (2014) *Central bank macroeconomic forecasting during the global financial crisis: The European Central Bank and Federal Reserve Bank of New York experiences.* **Journal of Business and Economic Statistics**, 32(4), 483-500.

[3] Andersson, K. M., Aranki, T., and Reslow, A. (2017) *Adjusting for information content when comparing forecast performance.* **Journal of Forecasting**, 36(7), 784-794.

[4] Andrews, D. W. K. (1991) *Heteroskedasticity and autocorrelation consistent covariance matrix estimation.* **Econometrica**, 59(3), 817-858.

[5] Bai, J., and Ng, S. (2005) *Tests for skewness, kurtosis, and normality for time series data.* **Journal of Business and Economic Statistics**, 23(1), 49-60.

[6] Blaskowitz, O., and Herwartz, H. (2014) *Testing the value of directional forecasts in the presence of serial correlation.* **International Journal of Forecasting**, 30(1), 30-42.

[7] Bobeica, E., and Jarociński, M. (2017) *Missing disinflation and missing inflation: the puzzles that aren't* . **ECB Working Paper Series**, 2000.

[8] BoE-IEO (2015) *Evaluating forecast performance.* **Bank of England - Independent Evaluation Office**.

[9] Bontemps, C., and Meddahi, N. (2005) *Testing normality: A GMM approach.* **Journal of Econometrics**, 124(1), 149-186.

[10] Champagne, J., Bellisle, G., and Sekkel R. (2018) *Evaluating the Bank of Canada staff economic projections using a new database of real-time data and forecasts.* **Bank of Canada Staff Working Paper**, 2018-52.

[11] Ciccarelli, M., and Osbat, C. (2017) *Low inflation in the euro area: Causes and consequences.* **ECB Occasional Paper Series**, 181.

[12] Clark, T. E., and McCracken, M. W. (2001) *Tests of equal forecast accuracy and encompassing for nested models.* **Journal of Econometrics**, 105(1), 85-110.

[13] Clark, T. E., and McCracken, M. W. (2017) *Tests of predictive ability for vector autoregressions used for conditional forecasting.* **Journal of Applied Econometrics**, 32(3), 533-553.

[14] Clements, M. P., Joutz, F. and Stekler, H. O. (2007) *An evaluation of the forecasts of the Federal Reserve: a pooled approach.* **Journal of Applied Econometrics**, 22(1), 121-136.

[15] Constâncio, V. (2015) *Understanding in inflation dynamics and monetary policy.* **Speech at the Jackson Hole Economic Policy Symposium**, 29 August 2015.

[16] Cumby, R. E., and Huizinga, J. (1992) *Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions.* **Econometrica**, 60(1), 185-195.

[17] Darvas, Z. (2018) *Forecast errors and monetary policy normalisation in the euro area.* **Bruegel, Policy Contribution**, December 2018, No. 24.

[18] Diebold, F. X., and Mariano, R. S. (1995) *Comparing predictive accuracy.* **Journal of Business and Economic Statistics**, 13(3), 253-263.

[19] Draghi, M. (2016a) *Introductory statement to the press conference with Q&A.* **Press conference**, 10 March 2016.

[20] Draghi, M. (2016b) *Delivering a symmetric mandate with asymmetric tools: monetary policy in a context of low interest rates.* **Speech at Oesterreichische Nationalbank**, 2 June 2016.

[21] ECB (2006) *Economic and monetary developments.* **ECB Monthly Bulletin**, June 2006, 75.

[22] ECB (2009) *New Procedure for constructing Eurosystem and ECB staff projection ranges.* **ECB publications-available online**, December 2009.

[23] ECB (2012) *The forecast bias of euro area HICP inflation.* **ECB Monthly Bulletin**, June 2012, 68-72.

[24] ECB (2013) *An assessment of Eurosystem staff macroeconomic projections.* **ECB Monthly Bulletin**, May 2013, 71-83.

[25] ECB (2014) *The ECB's forward guidance.* **ECB Monthly Bulletin**, April 2014.

[26] ECB (2016) *A guide to the Eurosystem/ECB staff macroeconomic projection exercises.* **ECB publications**, July 2016.

[27] Elliott, G., Komunjer, I., and Timmermann, A. (2005) *Estimation and testing of forecast rationality under flexible loss.* **Review of Economic Studies**, 72(4), 1107-1125.

[28] Elliott, G., and Timmermann, A. (2016) *Economic forecasting.* **Princeton University Press**.

[29] El-Shagi, M., Giesen, S. and Jung, A. (2016) *Revisiting the relative forecast performances of Fed staff and private forecasters: A dynamic approach.* **International Journal of Forecasting**, 32(2), 313-323.

[30] Faust, J., and Wright, J. H. (2008) *Efficient forecast tests for conditional policy forecasts.* **Journal of Econometrics**, 146, 293-303.

[31] Faust, J., and Wright, J. H. (2009) *Comparing Greenbook and reduced form forecasts using a large real-time dataset.* **Journal of Business & Economic Statistics**, 27(4), 468-479.

[32] Fioramanti, M., Gonzàlez Cabanillas, L., Roelstraete, B., and Ferrandis Vallterra, S. A. (2016) *European Commission's forecast accuracy revisited: Statistical properties and possible causes of forecast errors.* **European Economy - Discussion Papers**, 027.

[33] Freedman, C. (2014) *An evaluation of commissioned studies assessing the accuracy of IMF forecasts.* **IMF - Independent Evaluation Office Background Papers**, BP/14/02.

[34] Gavin, W. T. and Mandal, R. J. (2003) *Evaluating FOMC forecasts.* **International Journal of Forecasting**, 19(4), 655-667.

[35] Giacomini, R. and Rossi, B. (2010) *Forecast comparisons in unstable environments.* **Journal of Applied Econometrics**, 25(4), 595-620.

[36] Groen, J. J. J., Kapetanios, G., and Price, S. (2009) *A real time evaluation of Bank of England forecasts of inflation and growth.* **International Journal of Forecasting**, 25(1), 74-80.

[37] Harvey, D. I., Leybourne, S., and Newbold, P. (1997) *Testing the equality of prediction mean squared errors.* **International Journal of Forecasting**, 13(2), 281-289.

[38] Harvey, D. I., Leybourne, S., and Newbold, P. (1998) *Tests for forecast encompassing.* **Journal of Business and Economic Statistics**, 16(2), 254-259.

[39] Harvey, D. I., and Newbold, P. (2003) *The non-normality of some macroeconomic forecast errors.* **International Journal of Forecasting**, 19(4), 635-653.

[40] Holden, K., and Peel, D. A. (1990) *On testing for unbiasedness and efficiency of forecasts.* **The Manchester School**, 58(2), 120-127.

[41] IMF-IEO (2014) *IMF forecasts: Process, quality and country perspectives.* **IMF - Independent Evaluation Office**.

[42] Jarque, C. M., and Bera, A. K. (1987) *A test for normality of observations and regression residuals.* **International Statistical Review**, 55(2), 163-172.

[43] Kapetanios, G., Labhard, V., and Price, S. (2008) *Forecast combination and the Bank of England's suite of statistical forecasting models.* **Economic Modelling**, 25(4), 772-792.

[44] Ljung, G. M., and Box, P. E. G. (1978) *On a measure of a lack of fit in time series models.* **Biometrica**, 65(2), 297-303.

[45] Lobato, I. N., and Velasco, C. (2004) *A simple test of normality for time series.* **Econometric Theory**, 20(4), 671-689.

[46] Marcellino, M. (2000) *Forecast bias and MSFE encompassing.* **Oxford Bulletin of Economics and Statistics**, 62(4), 533-542.

[47] McCracken, M. W. (2007) *Asymptotics for out of sample tests of Granger causality.* **Journal of Econometrics**, 140(2), 719-752.

[48] Melander, A., Sismanidis, G., and Grenouilleau, D. (2007) *The track record of the Commission's forecasts - an update.* **European Economy - Economic Papers**, 291.

[49] Mincer, J. A., and Zarnowitz, V. (1969) *The evaluation of economic forecasts.* **NBER economic forecasts and expectations: Analysis of forecasting behaviour and performance**, 1-46.

[50] Newey, W., and West, K. (1987) *A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.* **Econometrica**, 55(3), 703-708.

[51] Newey, W., and West, K. (1994) *Automatic lag selection in covariance matrix estimation.* **The Review of Economic Studies**, 61(4), 631-653.

[52] Pain. N., Lewis, C., Dang, T. T., Jin, Y., and Richardson, P. (2014) *OECD forecasts during and after the financial crisis: A post mortem.* **OECD Economics Department Working Papers**, No. 1107.

[53] Paloviita, M., Haavio M., Jalasjoki P., and Kilponen, J. (2017) *What does "below, but close to, two percent" mean? Assessing the ECB's reaction function with real time data.* **Bank of Finland Research Discussion Papers**, 29/2017.

[54] Patton, A. J., and Timmermann, A. (2007a) *Properties of optimal forecasts under asymmetric loss and nonlinearity.* **Journal of Econometrics**, 140(2), 884-918.

[55] Patton, A. J., and Timmermann, A (2007b) *Testing forecast optimality under unknown loss.* **Journal of the American Statistical Association**, 102(480), 1172-1184.

[56] Patton, A. J., and Timmermann, A (2012) *Forecast rationality tests based on multi-horizon bounds.* **Journal of Business and Economic Statistics**, 30(1), 1-17.

[57] Pesaran, H. M., and Timmermann, A. (1992) *A simple non-parametric test of predictive performance.* **Journal of Business and Economic Statistics**, 10(4), 461-465.

[58] Pesaran, H. M., and Timmermann, A. (2009) *Testing dependence among serially correlated multicategory variables.* **Journal of the American Statistical Association**, 104(485), 325-337.

[59] Politis, D. N., and Romano, J. P. (1994) *The stationary bootstrap.* **Journal of the American Statistical Association**, 89(428), 1303-1313.

[60] Politis, D. N., and White, H. (2004) *Automatic block-length selection for the dependent bootstrap.* **Econometric Reviews**, 23(1), 53-70.

[61] Psaradakis, Z., and Vàvra, M. (2018) *Normality tests for dependent data: large-sample and bootstrap approaches.* **Communications in Statistics - Simulation and Computation**, 1-22.

[62] Reischneider, D., and Tulip, P. (2007) *Gauging the uncertainty of the economic outlook from historical forecasting errors.* **Finance and Economics Discussion Series, Federal Reserve Board**, 60.

[63] Reischneider, D., and Tulip, P. (2018) *Gauging the uncertainty of the economic outlook using historical forecasting errors: The Federal Reserve's approach.* **International Journal of Forecasting**.

[64] Romer, C. D., and Romer, D. H. (2000) *Federal Reserve information and the behaviour of interest rates.* **American Economic Review**, 90(3), 429-457.

[65] Tulip, P. (2006) *Has the economy become more predictable? Changes in Greenbook forecast accuracy.* **Journal of Money, Credit and Banking**, 41(6), 1217-1231.

[66] Tulip, P., and Wallace, S. (2012) *Estimates of uncertainty around the RBA's forecasts.* **Research Discussion Paper, Reserve Bank of Australia**, 7.

[67] Vogel, L. (2007) *How do the OECD growth projections for the G7 economies perform?: A post-mortem.* **OECD Economics Department Working Paper**, 573.

[68] West, K. D. (1996) *Asymptotic inference about predictive ability.* **Econometrica**, 64(5), 1067-1084.

[69] West, K. D., and McCracken, M. W. (1998) *Regression based tests of predictive ability.* **International Economic Review**, 39(4), 817-840.

Acknowledgements

We thank A. Page, N. Vidalis, M. Lenza, an anonymous referee and seminar participants in Frankfurt for useful comments and discussions. All remaining errors and omissions are ours. The views expressed in this article are those of the authors and do not necessarily reflect those of the European Central Bank or the Eurosystem.

Georgios Kontogeorgos
European Central Bank, Frankfurt am Main, Germany; email: georgios.kontogeorgos@ecb.europa.eu

Kyriacos Lambrias
European Central Bank, Frankfurt am Main, Germany; email: kyriacos.lambrias@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

PDF ISBN 978-92-899-3553-1 ISSN 1725-2806 doi:10.2866/518880 QB-AR-19-072-EN-N