How far can we forecast?

Statistical tests of the predictive content

Jörg Breitung* Malte Knüppel†

University of Cologne Deutsche Bundesbank

August 2, 2020

Abstract

We develop tests for the null hypothesis that forecasts become uninformative beyond some maximum forecast horizon h^*. The forecast may result from a survey of forecasters or from an estimated parametric model. The first class of tests compares the mean-squared prediction error of the forecast to the variance of the evaluation sample, whereas the second class of tests compares it to the mean-squared prediction error of the recursive mean. We show that the forecast comparison may easily be performed by adopting the encompassing principle, which results in simple regression tests with standard asymptotic inference. Our tests are applied to forecasts of macroeconomic key variables from the survey of Consensus Economics. The results suggest that these forecasts are barely informative beyond 2–4 quarters ahead.

Keywords: Hypothesis Testing, Forecast Evaluation, Forecast Horizon

JEL classification: C12, C32, C53.

*Corresponding author: University of Cologne, Institute of Econometrics, 50923 Cologne, Germany. Email: breitung@statistik.uni-koeln.de, malte.knueppel@bundesbank.de.

†The views expressed in this paper do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

1 Introduction

The choice of the largest forecast horizon appears to be an important issue for decision-makers. For example, in recent years, several central banks, including the Federal Reserve, the Bank of England and the European Central Bank (ECB), decided to increase the time horizons of their macroeconomic forecasts or surveys they conduct among private sector forecasters. For instance, since 2004, the Bank of England has published macroeconomic forecasts for up to 12 instead of 8 quarters ahead. In the US Survey of Professional Forecasters (SPF), conducted by the Federal Reserve Bank of Philadelphia, the largest horizon for forecasts of some annual variables like real GDP was extended from 1 to 3 years in 2010. In 2013 and 2014, the largest horizon of several macroeconomic forecasts by the ECB and in the ECB's SPF increased from 1 year to 2 years.[1] Yet, it is unclear whether forecasts for larger horizons actually provide valuable information in such cases, as forecast error variances approach the unconditional variance of the target variable.

For assessing the predictive content, Theil (1958) proposed (among other measures) the inequality coefficient that compares the actual forecast to some "naive" guess. If the forecast is informative, the inequality coefficient should be substantially smaller than unity, see e.g. Isiklar and Lahiri (2007) for an application to survey forecasts from Consensus Economics.

Using the unconditional mean as the uninformative benchmark, the inequality coefficient is related to the R^2 from a regression of the actual observations on their forecasts (often referred to as Mincer-Zarnowitz regression, see Mincer and Zarnowitz 1969). This R^2 was considered by Nelson (1976) and Diebold and Kilian (2001) as a measure for the predictive content. Diebold and Kilian (2001) generalized this measure to accommodate nonstationary time series and arbitrary loss functions. Their measure compares the loss of the short-run forecast to the loss of the long-run prediction. If the target variable is stationary, the loss function is quadratic, and the horizon of the long-run forecast tends to infinity, then the Diebold-Kilian measure and Nelson's R^2 coincide.

The empirical literature reports few and differing results concerning the largest informative forecast horizon. The differences are at least partly due to different data transformations, as pointed out by Galbraith and Tkacz (2007). For example, concerning quarterly GDP, they find

[1] See Knüppel (2018) for further details.

that forecasts of quarter-on-quarter growth are barely informative beyond a forecast horizon of one quarter. For year-on-year forecasts this horizon increases to about four quarters, which may not be surprising provided the overlap of the forecasts. Concerning annual GDP growth, Isiklar and Lahiri (2007) find that forecasts are informative for horizons up to six quarters. Diebold and Kilian (2001) report even larger horizons for HP-filtered or linearly detrended GDP.

The purpose of this paper is to provide statistical tests for assessing the predictive content of forecasts, thereby determining the largest informative forecast horizon. A natural way of testing is to compare the forecast to some uninformative benchmark. To this end, traditional forecast evaluation tests, such as the Diebold and Mariano (1995) test or forecast encompassing tests (e.g. Harvey, Leybourne, and Newbold 1997) can be adopted.[2] It is important to note that the uninformative benchmark is typically nested within the forecast under scrutiny in the sense that under the null hypothesis the difference between the forecasts tends to zero in probability (Clark and McCracken 2001). In this paper we propose an alternative approach that sidesteps the problem of selecting a "naive benchmark" and directly compares the mean-squared forecast error to the unconditional variance of the target variable.

We consider three different forecast scenarios. Whenever the forecasts are based on survey expectations (as in our empirical application), it is natural to assume that the forecasts correspond to some conditional mean based on an associate information set. We argue that it makes a crucial difference whether the forecasts are exactly identical to some conditional mean function (scenario 1) or whether the forecasts involve some additional noise (scenario 2). For the latter scenario it is natural to test the hypothesis of uninformative forecasts by running a Mincer-Zarnowitz regression, whereas under the first scenario the Mincer-Zarnowitz regression is invalid due to the fact that under the null hypothesis the conditional mean is constant. Yet, it is possible to construct a Diebold-Mariano type test for scenario 1 by taking into account the fact that the forecast comparison is nested. Scenario 3 assumes that the forecast is generated by an estimated model. In many cases the forecast results from a parametric specification of some conditional mean function, where the parameters are estimated from past observations. This scenario is related to scenario 2, where the noise corresponds to the estimation error. An important difference is, however, that the estimation error vanishes as the number of observations

[2]See Elliott and Timmermann (2016, chap. 17) and Cheng, Swanson, and Yao (2020) for reviews of the recent literature.

tends to infinity.

We consider two testing strategies for assessing the predictive power of the forecasts. As our test procedures are based on a comparison of the forecast and the unconditional mean as an uninformative benchmark, we require an estimator for the unconditional mean. One approach is to employ the in-sample mean of the evaluation sample. Alternatively, we may use some other uninformative benchmark such as the recursive mean computed from an expanding sample. It turns out that the in-sample version of the test typically provides a simpler test with less assumptions and choices to make. Moreover the in-sample version of our tests tend to perform better in many situations.

The rest of this paper is organized as follows. In Section 2 we introduce our testing framework. Tests of the information content of survey expectations are considered in Section 3, whereas Section 4 deals with forecasts based on parametric models. Section 5 investigates the small sample properties of the tests by means of Monte Carlo experiments, and in Section 6 the proposed tests are applied to forecasts of key macroeconomic variables as reported by Consensus Economics. Section 7 concludes. Additional results are provided in an online appendix to this paper.

2 Testing framework

Assume that the target time series $\{Y_t\}$ is generated by a stationary and ergodic stochastic process. The h-step ahead forecast of Y_{t+h} based on information up to time period t is denoted by $\widehat{Y}_{t+h|t}$. Under quadratic loss the optimal forecast equals the conditional expectation $\mu_{h,t} = \mathbb{E}(Y_{t+h}|\mathcal{I}_t)$, where $\mathcal{I}_t$ represents the information set at time period t. For our analysis we distinguish two time spans. The evaluation period starts at $t = 1 + h$ and runs up to period $t = n + h$. For these time periods we compare the forecasts $\widehat{Y}_{1+h|1}, \ldots, \widehat{Y}_{n+h|n}$ to the actual values $Y_{1+h}, \ldots, Y_{n+h}$. The following assumption characterizes the process generating the series to be forecasted by the information set $\mathcal{I}_t$:

ASSUMPTION 1 *(i) For each $h = 1, 2, \ldots$ the time series is decomposed as*

$$Y_{t+h} = \mu_{h,t} + u_{h,t} \ , \tag{1}$$

where $\mu_{h,t} = \mathbb{E}(Y_{t+h}|\mathcal{I}_t)$ *for* $h > 0$, $\mathcal{I}_t$ *denotes an increasing sigma-field and* $u_{h,t} = \phi_h(L)\varepsilon_{h,t}$, *where* $\phi_h(L) = 1 + \phi_{h,1}L + \phi_{h,2}L^2 + \cdots$ *is a lag polynomial with all roots outside the unit circle,* $\sum_{i=1}^{\infty} |\phi_{h,i}| < \infty$ *and* $\varepsilon_{h,t}$ *is an i.i.d. white noise process with* $\mathbb{E}(\varepsilon_{h,t}) = 0$ *and* $\mathbb{E}(\varepsilon_{h,t}^2) = \sigma_h^2$. *(ii)* $\mathbb{E}|u_{h,t}|^{2+\delta} < C < \infty$ *for some* $\delta > 0$. *(iii)* $\mathbb{E}\left(n^{-1}\sum_{t=1}^{n}\mu_{h,t}^2\right) < C < \infty$ *for all* n.

For some of our results, the assumptions of a linear process with constant variances are not necessary (see Remark 2 below) and may be relaxed at the cost of a more demanding notation and asymptotic analysis.

Let $\mu = \mathbb{E}(Y_t)$ denote the unconditional mean. We are interested in testing the null hypothesis

$$\textit{no information:} \quad \mathbb{E}(Y_{t+h} - \widehat{Y}_{t+h|t})^2 \geq \mathbb{E}(Y_{t+h} - \mu)^2, \text{ for } h > h^* \text{ and } t \in \{1, \ldots, n\}, \tag{2}$$

which is tested against the alternative $H_1 : \mathbb{E}(Y_{t+h} - \widehat{Y}_{t+h|t})^2 < \mathbb{E}(Y_{t+h} - \mu)^2$. The null hypothesis (2) asserts that there exists a maximum forecast horizon h^* beyond which the process Y_t is unpredictable with respect to the information set $\mathcal{I}_t$. If the forecast Y_{t+h} is identical to the conditional mean $\mu_{h,t}$, then the hypothesis (2) is equivalent to the hypothesis:

$$\textit{constant mean:} \quad \mathbb{E}(Y_{t+h}|\mathcal{I}_t) = \mu_{h,t} = \mu, \text{ for } h > h^* \text{ and } t \in \{1, \ldots, n\}, \tag{3}$$

that is, the conditional expectation is constant within the evaluation sample.

In many practical situations it is not reasonable to assume that the forecast is identical to some conditional expectation. In Section 3 we assume that the conditional expectation may be contaminated by some noise η_t such that $\widehat{Y}_{t+h|t} = \mu_{h,t} + \eta_t$. Another possibility is that the conditional expectation is specified as a parametric function involving a parameter vector θ, which needs to be estimated (see Section 4). In such cases the null hypotheses (2) and (3) are not equivalent as $\mathbb{E}(Y_{t+h} - \widehat{Y}_{t+h|t})^2$ may be larger than $\mathbb{E}(Y_{t+h} - \mu)^2$ due to the variance of the noise or the estimation error. Therefore, a test for a constant conditional mean may reject while a test of the hypothesis (2) is not able to reject the *no information* hypothesis. In Section 3 we show that if the conditional mean is contaminated with noise, then the *constant mean* hypothesis (3) implies that the slope coefficient of the Mincer-Zarnowitz regression is equal to zero, whereas

the *no information* hypothesis (2) refers to a slope not larger than 0.5. Whenever the forecast $\widehat{Y}_{t+h|t}$ converges in probability to the conditional expectation $\mu_{h,t}$ (scenario 3 for model-based predictions), then the *no information* hypothesis is asymptotically equivalent to the *constant mean* hypothesis.

Another difficulty with hypothesis (2) is that μ is not observed and has to be replaced by some estimate. Our preferred approach is to insert the mean of the evaluation sample $\overline{Y}^h = n^{-1} \sum_{t=1}^{n} Y_{t+h}$. Another possibility is to replace μ by an uninformative benchmark $\widehat{Y}_t^*$ known in period t, such as the recursive mean computed from observations prior to t or the mean of a rolling window. The advantage of employing a recursive mean is that the estimation error of $\widehat{Y}_t^* - \mu$ tends to become smaller as t increases. The mean of a rolling window is suitable for adopting the finite-sample framework of Giacomini and White (2006), see Section 5.

Using the uninformative benchmark $\widehat{Y}_t^*$ instead of the mean of the evaluation sample $\overline{Y}^h$ requires more information (a longer history of the target variable), a stronger assumption (the null hypothesis applies to a longer time span involving the risk of structural breaks) and additional choices (recursive vs. rolling mean, the choice of the estimation window size) to perform the tests.[3] Therefore, the tests based on $\overline{Y}^h$ are more versatile and can easily be employed, for example, when analyzing survey forecasts or comparing different forecasts (for instance, survey vs. model, model with larger estimation sample vs. model with shorter estimation sample etc.). However, if a forecaster is interested in the largest informative horizon of her model only, $\widehat{Y}_t^*$ and $\overline{Y}^h$ are both suitable choices, since the additional requirements related to $\widehat{Y}_t^*$ also apply to the model-based forecast anyway.[4]

The maximum forecast horizon h^* can be identified by sequentially applying a consistent test for horizons $h = 1, 2, ...$ until it is not rejected for the first time. Then, h^* is identified as the penultimate horizon tested. Provided that the tests are consistent, h^* is correctly identified with probability approaching $1 - \alpha$ as $n \to \infty$, where α denotes the significance level of the test. Therefore α must tend to zero to achieve a consistent selection rule for h^* (see Remark 6 below). It should be noted that the forecast error variances are monotonically increasing with respect

[3] Moreover, since the uninformative benchmark is typically nested within the (potentially) informative forecast, the "standard" Diebold and Mariano (1995) or encompassing tests are invalid.

[4] This case corresponds to scenario 3 and can be addressed using existing tests like the one proposed in Clark and West (2007). Yet, to the best of our knowledge, these tests have never been applied sequentially to find the largest informative horizon, as suggested in what follows.

to the forecast horizon (see, for instance, Patton and Timmermann 2012, Section 2.2). Thus, if a forecast is uninformative at some horizon h, it must also be uninformative for any higher horizon. Therefore, we can stop the testing sequence as soon as the test does not reject for the first time.

It is important to notice that there may not exist a finite maximum forecast horizon h^*. If, for example, Y_{t+h} is generated by an AR(1) process, then h^* is infinity. In such cases our tests address the question: "For how many time periods ahead does the forecast significantly outperform the naive benchmark?" As for many other statistical tests, failing to reject the null hypothesis does not imply that it is true.

3 Survey expectations

First we focus on forecasts that are not based on an (estimated) statistical model but result from expectations of a sample of individuals. We consider two different scenarios: In scenario 1, the expectation is identical to some conditional mean, that is, $\widehat{Y}_{t+h|t} = \mu_{h,t} = \mathbb{E}(Y_{t+h}|\mathcal{I}_t)$. For our test it is not important to specify and know the information set $\mathcal{I}_t$. It is only required that there exists some sequence of increasing information sets with $\mathcal{I}_t \in \mathcal{I}_{t+1}$. In scenario 2 the conditional expectation is observed with noise such that $\widehat{Y}_{t+h|t} = \mu_{h,t} + \eta_t$. The error term η_t may be due to reporting error or forecast disagreement, for instance.[5]

3.1 Tests without expectation error

First we analyse scenario 1 where the survey expectations are identical to the conditional expectation based on some information set $\mathcal{I}_t$. In this setup the *no information* hypothesis (2) and the *constant mean* hypothesis (3) are equivalent. To test the null hypothesis, the unknown unconditional mean μ may be replaced by the in-sample mean $\overline{Y}^h = n^{-1}\sum_{t=1}^{n} Y_{t+h}$. Another alternative is to employ some uninformative benchmark such as the recursive mean based on

[5] In the literature on survey expectations (e.g. Carlson and Parkin 1975) it is often assumed that individual expectations are drawn from the distribution $\mathcal{N}(\mu_{h,t}, \sigma_h^2)$. If m_t is the number of survey participants in period t, the error of the survey mean is distributed as $\eta_t \sim \mathcal{N}(0, \sigma_h^2/m_t)$.

$T+t$ observations:

$$\overline{Y}_t = \frac{1}{T+t} \sum_{s=-T+1}^{t} Y_s \ . \tag{4}$$

The test statistics are based on the mean-squared prediction error (MSPE) loss differentials:

$$\delta^h_{0,t} = u^2_{h,t} - (Y_{t+h} - \overline{Y}^h)^2 \tag{5}$$

$$\delta^h_{T,t} = u^2_{h,t} - (Y_{t+h} - \overline{Y}_t)^2 \ , \tag{6}$$

where $u_{h,t} = Y_{t+h} - \widehat{Y}_{t+h|t}$. Following Diebold and Mariano (1995, henceforth DM) we construct two test statistics based on $\delta^h_{0,t}$ and $\delta^h_{T,t}$. Notice that the forecast comparison $\delta^h_{T,t}$ is based on nested forecasts (see Clark and McCracken 2001) implying that under the null hypothesis $\delta^h_{0,t} \xrightarrow{p} 0$ as $n \to \infty$ and $\delta^h_{T,t} \xrightarrow{p} 0$ as $T \to \infty$.

THEOREM 1 *The DM type test statistics are defined as*

$$dm_{0,h} \ = \ \frac{1}{\omega^2_h} \sum_{t=1}^{n} \delta^h_{0,t} \quad \textit{and} \quad dm_{T,h} = \frac{1}{\omega^2_h} \sum_{t=1}^{n} \delta^h_{T,t} \tag{7}$$

where

$$\omega^2_h = \lim_{n\to\infty} \mathbb{E} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_{h,t} \right)^2 . \tag{8}$$

Under the null hypothesis $H_0 : \mu_{h,t} = \mu$ *for all* t *and* $h > h^*$, *Assumption 1, a recursive forecasting scheme with* $T \to \infty$, $n \to \infty$ *and* $T/(T+n) \to \pi \in [0,1)$ *the test statistics are distributed as*

$$dm_{0,h} \xrightarrow{d} \chi^2 \tag{9}$$

$$dm_{T,h} \xrightarrow{d} 2 \int_\pi^1 \frac{1}{a} W(a)\, dW(a) - \int_\pi^1 \frac{1}{a^2} W(a)^2 da. \tag{10}$$

where $W(a)$ *represents a standard Brownian motion defined on* $[0,1]$.

REMARK 1 The test statistics (7) reveal some interesting differences to the original DM statistic. First, the sum of the loss differential is not divided by $\sqrt{n}$. Second, the statistics involve the long-run variance of $u_{h,t}$ instead of the square root of the long-run variance of the loss differentials. Third, the limiting distribution is different from a standard normal distribution. This is due to the nested nature of the forecast comparison. It is important to notice that for the test based on the recursive mean, $dm_{T,h}$, the limiting distribution depends on the fraction π. Critical values for selected values of π are presented in the online appendix. In contrast, the limiting distribution of the in-sample statistic $dm_{0,h}$ does not depend on π and is available from standard statistical tables and software. Note that the critical values are obtained from the *lower* quantiles of the χ^2 distribution. For example, the critical value for a significance level of 0.05 is 0.0039.

REMARK 2 Following Diebold and Mariano (1995), the long-run variance ω_h^2 can be estimated as

$$\widehat{\omega}_h^2 = \frac{1}{n}\sum_{t=1}^{n} u_{h,t}^2 + \frac{2}{n}\sum_{j=1}^{h-1}\sum_{t=j+1}^{n} u_{h,t}u_{h,t-j}$$

It should be noted, however, that by applying a rectangular kernel, the estimated long-run variance may be negative. In this case some other kernel should be applied that ensures a positive estimator for the long-run variance (e.g. Newey and West 1987). Note also that the usual estimators for the long-run variance are robust to heteroskedasticity. Accordingly, Assumption 1 may be generalized to allow for heteroskedastic processes when the statistic $dm_{0,h}$ is concerned. On the other hand, the limiting distribution of the test statistic $dm_{T,h}$ depends on functionals of Brownian motions that are affected whenever $u_{h,t}$ is heteroskedastic.

REMARK 3 The statistic $dm_{T,h}$ employs T additional observations prior to the evaluation sample, requiring the assumption that the unconditional mean remains constant during the entire time span of $T+n$ time periods. In contrast, the statistic $dm_{0,h}$ is less vulnerable to structural instability. Surprisingly, using more information does not imply that the statistic $dm_{T,h}$ is more powerful than $dm_{0,h}$, as documented in Section 5.

3.2 Tests with expectation error

Let us now move on to scenario 2 where the forecast is contaminated with noise, i.e. $\widehat{Y}_{t+h|t} = \mathbb{E}(Y_{t+h|t}|\mathcal{I}_t) + \eta_t$ and η_t represents the noise. Our asymptotic analysis is based on the following assumption:

ASSUMPTION 2 *(i) The noise η_t is generated by a stationary process with $\mathbb{E}(\eta_t) = 0$, $\mathbb{E}(\eta_t^2) = \sigma_\eta^2$, $\mathbb{E}(\eta_t^4) < \infty$ and long-run variance $\omega_\eta^2 = \lim_{n\to\infty} \mathbb{E}\left[n^{-1}\sum_{t=1}^n \eta_t\right]^2 < \infty$. (ii) $\mathbb{E}(\eta_t \mu_{h,t}) = 0$ for all t. (iii) Let $\xi_t = \eta_t u_{h,t}$. For $h > h^*$ and all t we assume that $\mathbb{E}(\xi_t) = 0$, $E(\xi_{t-j}\xi_t) = 0$ for $|j| \geq h$, and $\mathbb{E}|\xi_t|^{2+\delta} < \infty$ for some $\delta > 0$.*

Again, the assumption that the expectation error is homoskedastic is made to facilitate the proofs but is not necessary as the test statistic employs heteroskedasticity- and autocorrelation-consistent (henceforth HAC) standard errors. Assumptions 2 (ii) and (iii) ensure that the noise does not result in a systematic bias. The limitation to autocorrelation up to $h-1$ lags in (iii) is due to the fact that the HAC t-statistic assumes an MA$(h-1)$ process for $u_{h,t}$. We can easily relax this assumption to allow for some higher order correlation of ξ_t by employing a larger truncation lag for the HAC correction.

Our test of the *no information* hypothesis (2) relies on the following lemma:

LEMMA 1 *Let $\widehat{Y}_{t+h|t} = \mu_{h,t} + \eta_t$. (i) The* no information *hypothesis (2) and Assumptions 1 – 2 imply $\beta_h = 0.5$ in the regression:*

$$Y_{t+h} = \alpha_h + \beta_h \widehat{Y}_{t+h|t} + v_{t+h} \ . \tag{11}$$

To provide an intuitive explanation for this result we note that regression (11) is asymptotically equivalent[6] to running the forecast encompassing regression (cf. Elliott and Timmermann 2016,

[6]The only difference is that in (11) the implicit centering of the regressor is around $\overline{\widehat{Y}}_h = n^{-1}\sum_{t=1}^n \widehat{Y}_{t+h|t}$, whereas the regressor in (12) is centered around $\overline{Y}^h$. Assumptions 1 – 2 imply that $\overline{Y}^h \xrightarrow{p} \mu$ and $\overline{\widehat{Y}}_h \xrightarrow{p} \mu$ as $n \to \infty$.

pp. 393-397)

$$\begin{aligned} Y_{t+h} &= (1-\beta_h)\overline{Y}^h + \beta_h \widehat{Y}_{t+h|t} + \tilde{v}_{t+h} \\ \text{or } Y_{t+h} - \overline{Y}^h &= \beta_h(\widehat{Y}_{t+h|t} - \overline{Y}^h) + \tilde{v}_{t+h} \ . \end{aligned} \tag{12}$$

The representation (12) implies that for the hypothesis $\beta_h = 0.5$ the optimal forecast combination attaches equal weights to both forecasts. In other words none of these two forecasts is favored under the null hypothesis. The relevant alternative assigns larger weight to the forecast $\widehat{Y}_{t+h|t}$ than implied by the respective null hypothesis. Therefore, we consider one-sided tests, that is, the null hypothesis $\beta_h = 0.5$ is tested against $\beta_h > 0.5$.

All tests of the parameters β_h (and γ_h, see Remark 4) rely on the LM version of the (HAC) t-statistic constructed as

$$\tau_a = \frac{1}{\widehat{\omega}_a \sqrt{n}} \sum_{t=1}^{n} a_t \tag{13}$$

where $\widehat{\omega}_a^2$ denotes some consistent estimator for the long-run variance of a_t. The specific form of the sequence a_t is given in Theorem 2. The test statistic τ_a is asymptotically equivalent to the ordinary HAC t-statistics of the coefficients β_h in regression (11). The only difference of the latter statistic is that for estimating the long-run variance of a_t, the hypothesized coefficient $\beta_h = 0.5$ is replaced by the estimated coefficient $\widehat{\beta}_h$. Since $\widehat{\beta}_h$ is a consistent estimator for β_h, both versions of the tests are asymptotically equivalent under the null hypothesis.

It is interesting to consider the null hypothesis $\beta_h = 0$ which refers to the *constant mean* hypothesis (3). This null hypothesis implies that the forecast $\widehat{Y}_{t+h|t}$ and the actual value Y_{t+h} are uncorrelated. Since Assumption 2 supposes that η_t and Y_{t+h} are uncorrelated, it follows from the null hypothesis that $\mu_{h,t}$ and Y_{t+h} are uncorrelated as well. Since $\mathbb{E}[(Y_{t+h} - \mu)(\mu_{h,t} - \mu)] = \mathbb{E}(\mu_{h,t} - \mu)^2$ we conclude that $\beta_h = 0$ implies $\mu_{h,t} = \mu$ and, therefore, the test of $\beta_h = 0$ is equivalent to testing the *constant mean* hypothesis (3). In other words, $\beta_h = 0$ makes a statement about the conditional mean $\mu_{h,t}$, whereas $\beta_h = 0.5$ tests the hypothesis that the MSPE of the forecast fails to be smaller than the unconditional variance.

Another implication of the null hypothesis $\beta_h = 0$ is that there exists no *linear transformation* of the forecast $\widehat{Y}_{t+h|t}$ that results in a smaller MSPE than the unconditional variance. If $0 <$

$\beta_h < 0.5$, then the MSPE of the forecast $\widehat{Y}_{t+h,t}$ is larger than the unconditional variance of Y_{t+h}, but the linear transformation $\widehat{Y}^*_{t+1|t} = \alpha_h + \beta_h \widehat{Y}_{t+1|t}$ is an informative forecast. This is due to the fact that whenever the R^2 of the regression (11) is larger than zero, then the residual variance is smaller than the variance of Y_{t+h}. Note that the estimated linear transformation of $\widehat{Y}^*_{t+h|t}$ can also be represented as a linear combination of $\widehat{Y}_{t+h|t}$ and the in-sample mean. Therefore, the test for $\beta_h = 0$ can be interpreted as a type of forecast encompassing test.

The following theorem summarizes the asymptotic distributions of the HAC t-statistics for the *no information* hypothesis (2) and the *constant mean* hypothesis (3) based on the in-sample mean $\overline{Y}^h$. The tests employing the recursive mean $\overline{Y}_t$ are considered in Remark 4 and Appendix A.

THEOREM 2 *(i) Under Assumptions 1 – 2, $h > h^*$, $\sigma^2_\eta > 0$ and $n \to \infty$ the HAC t-statistics constructed as in (13) with*

$$a_t = \left[Y_{t+h} - \overline{Y}^h - 0.5(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h)\right](\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h) \quad \textit{for } H_0 : \beta_h = 0.5$$
$$a_t = \left(Y_{t+h} - \overline{Y}^h\right)(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h) \quad \textit{for } H_0 : \beta_h = 0$$

in the regression (11) possess a limiting standard normal distribution, with $\overline{\widehat{Y}}_h = n^{-1}\sum_{t=1}^n \widehat{Y}_{t+h|t}$.

The proof is relegated to Appendix B.

REMARK 4 In Appendix A we consider analogous tests based on the recursive mean as uninformative benchmark. Essentially this version of the test replaces the constant μ by the recursive mean $\overline{Y}_t$ and employs the HAC t-statistic for the hypothesis $\gamma_h = 0.5$ or $\gamma_h = 0$ in the regression

$$Y_{t+h} - \overline{Y}_t = \gamma_h \left(\widehat{Y}_{t+h|t} - \overline{Y}_t\right) + \nu_{t+h}. \tag{14}$$

It is interesting to note that this test is related to the adjusted MSPE statistic suggested by Clark and West (2007) for nested forecast comparisons. Their statistic is given by

$$\text{MSPE-adj} = \frac{2}{n}\sum_{t=1}^{n}\left(Y_{t+h} - \overline{Y}_t\right)\left(\widehat{Y}_{t+h|t} - \overline{Y}_t\right) \tag{15}$$

which is essentially equal to the numerator of the OLS estimator of γ_h multiplied by the factor $2/n$. As argued by Clark and West (2007) the adjustment accounts "for the noise associated with the larger model's forecast", whereas in our framework the adjusted MSPE statistic is equivalent to testing the hypothesis $\gamma_h = 0$ instead of $\gamma_h = 0.5$. As argued above, testing $\gamma_h = 0$ corresponds to the hypothesis that there does not exist a linear transformation of the forecast $\widehat{Y}_{t+h|t}$ with a MSPE lower than the unconditional variance. Likewise, the linear transformation can be regarded as a linear combination of $\widehat{Y}_{t+h,t}$ and the recursive sample mean.

REMARK 5 The tests considered in Theorem 2 have two important characteristics. First, they are valid even if the survey expectations are biased such that $\mathbb{E}(\widehat{Y}_{t+h|t}) = \mu_{t+h,t} + \psi$, where the bias ψ is constant over time. For instance, the survey expectations may be biased due to an asymmetric loss function, but nevertheless informative in the sense that if the survey participants expect an increase, the actual value is likely to increase as well. The invariance to a possible bias is due to the fact that the regression constant takes into account any deviation between the means of the forecast and the target variable. This is an important difference to the test considered in Remark 4, where the uninformative benchmark is the recursive mean and the resulting test statistic is not invariant to a forecast bias. Second, the tests do not run into problems if the noise η_t is small, as the regressor is well behaved for all $\sigma_\eta^2 > 0$. The reason is that the test statistic is invariant to the scaling of the regressor. Under the null hypothesis, the long-run variance of the regressor $\widehat{Y}_{t+h|t}$ may become arbitrarily small as long as it remains positive. In contrast, the proof of Theorem A.2 reveals that the regression t-statistic for $\gamma_h = 0.5$ and $\gamma_h = 0$ in Remark 4 involves an extra term due to $\overline{Y}_t - \mu$ that becomes relatively more important the smaller σ_η^2 is. This additional term results in severe size distortions whenever σ_η^2 is small. In empirical practice, the long-run variance of η_t is unknown and, consequently, the effect of the additional term on the asymptotic distribution is not clear.

4 Model predictions

In scenario 3, we consider model-based forecasts that are characterized by a conditional mean function $\mathbb{E}(Y_{t+h}|\mathcal{I}_t) = \mu_{h,t}(\theta)$, where the $k \times 1$ vector θ represents the model parameters. To economize on notation we do not make explicit the dependence of the forecast model on addi-

tional variables. In practice the unknown parameter vector θ is replaced by consistent estimates $\widehat{\theta}_t$ based on the recursive sampling scheme $\{-T+1,\ldots,0,1,\ldots,t\}$. Accordingly the estimated conditional mean is denoted by $\widehat{Y}_{t+h|t} \equiv \mu_{h,t}(\widehat{\theta}_t)$. To some extent this framework is related to scenario 2 where the survey expectations are contaminated by noise, a situation that was analysed in the previous section. The crucial difference is, however, that the estimation error tends to zero as T tends to infinity, whereas the variance of the expectation error η_t is assumed to be constant. Specifically we make the following assumptions on the estimated forecast function:

ASSUMPTION 3 *(i) Under the null hypothesis there exists some h^* such that $\mu_{h,t}(\theta) = \mu$ for all $h > h^*$. (ii) The parameters are estimated consistently with*

$$\begin{aligned} a) \quad & \widehat{\theta}_0 - \theta \;=\; O_p(T^{-1/2}) \\ b) \quad & \sup_{t\in\{1,\ldots,n\}} ||\widehat{\theta}_t - \widehat{\theta}_0|| \;=\; O_p\left(\frac{\sqrt{n}}{T}\right) \quad \textit{for } t = 1,2,\ldots,n \end{aligned}$$

where $\widehat{\theta}_0$ denotes the estimator based on time periods $\{-T+1,\ldots,-1,0\}$.
(iii) Let $D_{h,t}(\theta) = \partial\mu_{h,t}(\theta)/\partial\theta$ and $\overline{D}_h(\theta) = n^{-1}\sum_{t=1}^{n} D_{h,t}(\theta)$. For all $\theta_i^* \in [\theta_i - \epsilon, \theta_i + \epsilon]$ with $\epsilon > 0$ and $\theta^* = (\theta_1^*,\ldots,\theta_k^*)'$ it holds that

$$\begin{aligned} & \frac{1}{n}\sum_{t=1}^{n} ||D_{h,t}(\theta^*) - \overline{D}_h(\theta^*)|| \;\xrightarrow{p}\; \overline{D} \quad \textit{with } 0 < \overline{D} < \infty \\ & \mathbb{E}||D_{h,t}(\theta^*)u_{h,t}||^{2+\delta} < \infty \quad \textit{for some } \delta > 0 \textit{ and all } t. \end{aligned}$$

Part *(i)* refers to the *constant mean* hypothesis (3). Since for all $t \in \{1,\ldots,n\}$ we have $\mu_{t+h|t}(\widehat{\theta}_t) - \mu_{t+h|t}(\theta) \xrightarrow{p} 0$ as $T \to \infty$, this null hypothesis is asymptotically equivalent to the *no information* hypothesis (2). Therefore we focus on the hypothesis $\beta_h = 0$ which results in more powerful tests than testing $\beta_h = 0.5$. Part *(ii) a)* supposes the usual (parametric) convergence rate of the estimation error in the estimated parameter vector $\widehat{\theta}_0$ based on the pre-evaluation sample $t \in \{-T+1,\ldots,0\}$, whereas *(ii) b)* limits the variation of estimators in the recursive estimation scheme within the evaluation sample. Assumption 3 (iii) ensures the existence of a central limit theorem.

For illustration, consider the forecast based on a simple regression model with $\widehat{Y}_{t+h|t} = \widehat{\beta}_t X_t$,

where $\widehat{\beta}_t$ is the least squares estimator based on the $T+t$ time periods $\{-T+1,\ldots,t\}$. If X_t is stationary, then $\widehat{\beta}_0 - \beta = O_p(T^{-1/2})$ and Assumption 2 *(ii) a)* is fulfilled. Furthermore, it is not difficult to show[7] that $\widehat{\beta}_t - \widehat{\beta}_0 = O_p\left(\sqrt{t}/T\right)$ and, because $t \leq n$, Assumption 3 (ii) *b)* is satisfied. Furthermore, $D_{h,t}(\theta) = X_t$ and, assuming stationary regressors with positive variance, Assumption 3 (iii) is fulfilled as well. It should also be noted that this assumption rules out forecasts based on non-parametric estimators that typically involve lower convergence rates. In such cases T must grow faster to achieve a similar accuracy of the asymptotic approximations.

In an earlier version of this paper (Breitung and Knüppel 2018) we analyzed the asymptotic properties of a DM type test $dm_{0,h}$ considered in Theorem 1 above. Specifically we showed that the estimation error of such a test can be ignored if $T \to \infty$, $n \to \infty$ and $n/T \to 0$. Unfortunately, in typical sample sizes the additional term due to the estimation error remains large relative to the critical value and, therefore, the size distortions are substantial and disappear very slowly with increasing T. We therefore do not consider the test statistics $dm_{0,h}$ or $dm_{T,h}$ in this section. Rather we focus on the regression variant by testing the hypothesis $\beta_h = 0$ in (11).

In our asymptotic analysis we first focus on the case that n/T tends to zero. Although in empirical practice n/T is often in the range $0.2 - 0.5$, say, the test performs nevertheless reasonably well, even for sizable values of n/T. Note that the test statistic has a standard normal limiting distribution regardless of the fraction n/T, if the forecast were computed from a fixed forecasting scheme, where the estimated parameter values from period $t = 0$ are used for estimating the conditional mean function. Since the difference $\widehat{\theta}_t - \widehat{\theta}_0$ is typically small (see Assumption 3 (ii) *b*) the difference between applying the recursive scheme involving $\widehat{\theta}_t$ and the fixed scheme $\widehat{\theta}_0$ is typically small if T is reasonably large. Denote the respective forecast based on the fixed forecasting scheme as $\widehat{Y}^0_{t+h|t} = \mu_{h,t}(\widehat{\theta}_0)$, where $\widehat{\theta}_0$ denotes the estimate of θ using information from $t \in \{-T+1,\ldots,0\}$. Under the null hypothesis Assumption 3 implies

$$\mathbb{E}\left[(Y_{t+h} - \overline{Y}^h)\widehat{Y}^0_{t+h|t}\right] = \mathbb{E}\left[(u_{h,t} - \overline{u}^h)\widehat{Y}^0_{t+h|t}\right] = 0$$

with $\overline{u}^h = n^{-1}\sum_{t=1}^n u_{h,t}$, and it follows that the HAC t-statistic of $\beta_h = 0$ in the regression (11) possesses a standard normal limiting distribution regardless of the estimation error in $\widehat{Y}^0_{t+h|t}$.

[7] See the working paper version of this paper, Breitung and Knüppel (2018).

This is due to the fact that the estimation error $\widehat{Y}^0_{t+h|t} - \mu_{h,t}(\theta)$ and $u_{h,t} - \overline{u}^h$ are uncorrelated. In a recursive forecasting scheme the difference between $\widehat{Y}_{t+h|t}$ and $\widehat{Y}^0_{t+h|t}$ introduces a correlation with $\overline{u}^h$ which is due to the overlap of information employed in $\widehat{Y}_{t+h|t}$ and $\overline{Y}^h$ (resp. $\overline{u}^h$). This correlation gives rise to a negative bias that disappears as $n/T \to 0$. In practice, this bias is relatively small and results in a test that tends to be slightly conservative for sizable values of n/T.[8] The details are provided in the proof of the following theorem:

THEOREM 3 *Under Assumptions 1 and 3, a recursive forecasting scheme, $h > h^*$, $T \to \infty$, $n \to \infty$ and $n/T \to 0$ the HAC t-statistic (13) for testing the hypothesis $\beta_h = 0$ with*

$$a_t = \left(Y_{t+h} - \overline{Y}^h\right)\left(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h\right)$$

in the regression (11) possesses a standard normal limiting distribution, with $\overline{\widehat{Y}}_h = n^{-1}\sum_{t=1}^n \widehat{Y}_{t+h|t}$.

The proof is relegated to the online appendix.

REMARK 6 As mentioned in Section 2, a consistent selection rule for the maximum forecast horizon h^* requires that the size of the test tends to zero as $n \to \infty$. One possibility is to apply a critical value of the form $\kappa \log(n)$ with some $\kappa > 0$. This choice is motivated by the Bayesian information criterion. It is not difficult to see that under the alternative $\tau_a = O_p(n^{1/2})$ such that for $h \leq h^*$ we obtain $\lim_{n\to\infty} P(\tau_a > \kappa \log(n)) = 1$, whereas for $h > h^*$ we have $\tau_a = O_p(1)$ and $\lim_{n\to\infty} P(\tau_a > \kappa \log(n)) = 0$. Thus the decision rule based on the last rejection in the sequence of tests with $h = 1, 2, \ldots$ is weakly consistent. For instance, letting $n = 27$ the critical value $\frac{1}{2}\log(27) = 1.65$ is similar to the one-sided 0.05 critical value of a standard normal distribution. This suggests setting $\kappa = 0.5$ in order to obtain a selection rule roughly equivalent to usual hypothesis testing when n is small. The same considerations apply to the tests considered in Section 3.

[8] Calhoun (2016) suggests a related approach for sidestepping the problems due to the overlap of the two forecasts. He considers the test statistic proposed by Clark and West (2007), where the model forecast is computed following a rolling window forecasting scheme, whereas the benchmark is computed recursively.

5 Small sample properties

To compare the small sample properties of the proposed test statistics in alternative forecasting scenarios, we conduct a number of Monte Carlo experiments. As the main conclusions are robust against variants of the forecasting model and the forecast horizon we focus on the data-generating process (DGP) given by $Y_t = a + bX_{t-1} + \varepsilon_t$, where ε_t and X_t are independent standard normal random variables. For $b = 0$ the time series is unpredictable at all forecast horizons h, whereas for $b \neq 0$ the forecast $\widehat{Y}_{t+1|t} = a + bX_t$ is informative. This forecast corresponds to scenario 1 in Section 3 where we assume that the forecast is identical to the conditional mean of the process given the information $\mathcal{I}_t = \{X_t, X_{t-1}, \ldots\}$. In scenario 2 we assume that the forecast from scenario 1 is contaminated by the noise term η_t, which is again an independently and normally distributed random variable with $\mathbb{E}(\eta_t) = 0$ and $E(\eta_t^2) = \sigma_\eta^2$.

Table 1 compares the actual sizes of the test procedures proposed in Theorems 1 and 2. The upper panel reports the rejection rates for $b = 0$, according to the *constant mean* hypothesis (3) for scenario 1, i.e. in the case of no noise ($\sigma_\eta = 0$). It turns out that the size of the tests considered in Theorem 1 is very accurate for all combinations of n and T. The second panel depicts the actual sizes for the tests if the forecasts are contaminated by noise. The findings suggest that the in-sample test $dm_0 \equiv dm_{0,1}$ is quite sensitive to the noise whereas the test $dm_T \equiv dm_{T,1}$ based on the recursive mean as the benchmark is more robust at least if the noise is as small as $\sigma_\eta = 0.01$. In contrast, the tests that allow for noise perform well if the noise is large, but the test for $\gamma \equiv \gamma_1 = 0$ reveals severe size distortions in the case with very small noise ($\sigma_\eta = 0.001$). This is due to the fact that with small noise, the test statistic is dominated by a term with a nonstandard distribution that is related to the recursive mean $\overline{Y}_t - \mu$. The test for $\beta \equiv \beta_1 = 0$, however, performs well even for very small noise. The same holds for the test for $\beta = 0.5$ in the situation where $b = \sigma_\eta$ holds, such that the noisy forecasts are as accurate as forecasts based on the unconditional mean and the *no information* hypothesis holds. The test for $\gamma = 0.5$ again suffers from pronounced size distortions unless σ_η is large. Yet, these distortions are not as severe as with the test for $\gamma = 0$ in the case with $b = 0$.

[Table 1 about here.]

To study the relative power of the tests we let $b = 0.2$. We only present results for the

empirically relevant scenario 2, i.e. for forecasts with noise, and we only consider cases where the tests show reliable size properties in Table 1. The results presented in the upper panel of Table 2 indicate that the in-sample version of the DM type test is more powerful than the test with the recursive mean as a benchmark whenever the fraction n/T is larger than 0.1. If T gets very large, however, the test dm_T outperforms the in-sample version dm_0. Compared to the tests of Theorem 2 (with noise) it turns out that the robustness to noise comes at the expense of a slight loss of power. Note that the null hypothesis of the tests of Theorem 1 implies $\mu_{1,t} = \mu$. This corresponds to the null hypothesis $\beta = 0$ in the tests of Theorem 2. For large n both tests have similar power, but for n as small as 25 or 50 the corresponding test in Theorem 2 ($\beta = 0$) is substantially less powerful.

[Table 2 about here.]

Let us now turn to scenario 3, i.e. to the tests for model-based forecasts. In our example, the forecasts are obtained as $\widehat{Y}_{t+1|t} = \widehat{a}_t + \widehat{b}_t X_t$, where $(\widehat{a}_t, \widehat{b}_t)$ refers to the OLS estimates based on time periods $\{-T+1, \ldots, t\}$. Our findings are summarized in Table 3. The results presented in the upper panel report the actual sizes of the tests. It turns out that the tests are conservative for all combinations of n and T. The size distortions appear to depend on the fraction n/T. For $n/T < 1$, as often encountered in practice, the size distortions tend to be rather small. For a nominal size of 0.05, the actual size is usually in the range 0.02 – 0.04.[9] The power of the tests is presented in the lower half of Table 3. Note that the test statistics for model-based forecasts are similar to the tests for survey forecasts with noise; the only difference is how the forecasts relate to the conditional expectation $\mu_{h,t} = \mathbb{E}(Y_{t+h}|X_t, X_{t-1}, \ldots)$. In Theorem 2 the noise is assumed to be uncorrelated with $\mu_{h,t}$ and the variance does not depend on n or T. In contrast, the size of the "noise" that is due to the estimation error is a function of n and T and it is not uncorrelated with $\mu_{h,t}$. To get an idea of the size of the estimation error we calculated the standard deviation of the estimated conditional mean function as 0.1 for $T = 100$ which corresponds to the case with large noise in Table 1. The important difference is, however, that the estimation error is not independent of the sample mean $\overline{Y}^h$, which tends to result in moderate size distortions unless n/T is very small.

[9] In simulations not reported here, we tried out values of n/T as large as 40, but the actual size never dropped below 0.01.

[Table 3 about here.]

We finally consider the test procedure of Giacomini and White (2006). This test involves a rolling window forecasting scheme with fixed window size B for the forecast model and the benchmark, and it simply tests if their loss differential denoted by $\widetilde{dm}_{B,1}$ equals zero.[10] The left panel of Table 4 reports the actual sizes of this test for various window sizes. As the test is based on the small sample comparison of the losses, the coefficient b is calibrated such that under the null hypothesis the expected losses are identical. The corresponding values of b are presented in the second column of Table 4, and the note contains details about their calibration. From the simulation results it turns out that for small evaluation samples ($n = 25$), the test is slightly oversized whereas for $n \geq 100$ the test appears to be slightly conservative. This corresponds to the results of McCracken (2019) who found that the Giacomini-White test tends to be conservative for large n. With respect to the power of the test, we observe that – as expected – the power of the test increases gradually with B. Compared to the encompassing tests proposed in Theorem 3 we observe a severe loss of power. For example, while the Giacomini-White test rejects in 23 percent of the cases if $n = 100$ and $B = 250$, the encompassing test based on γ rejects in at least 60 percent of the cases if $T \geq 250$ and $n = 100$.

[Table 4 about here.]

In the online appendix we present additional Monte Carlo experiments for multiple forecast horizons that by and large corroborate our results for $h = 1$.

6 Empirical results

For the empirical application of the tests, we employ quarterly survey forecasts collected by Consensus Economics. The mean of the forecasts across all panelists is known to be a very accurate forecast, as documented, for example, by Ang, Bekaert, and Wei (2007) for inflation forecasts. While survey forecasts are commonly evaluated ignoring the potential presence of any type of noise (see Clements 2019, chap 4.1), we are going to focus on the tests for scenario 2, i.e.

[10] This test is loosely related to scenario 2, because the parameter estimation error resulting from the rolling window is stationary and can be regarded as a form of autocorrelated noise. However, in contrast to all three scenarios considered, the uninformative benchmark is likewise contaminated by autocorrelated noise even as $n \to \infty$.

on forecasts with noise. Therefore, we employ the tests for $\beta_h = 0$ and $\gamma_h = 0$, and for $\beta_h = 0.5$ and $\gamma_h = 0.5$. We do so because disagreement across forecasters combined with entry and exit of forecasters inevitably leads to some form of noise.[11]

We consider quarterly forecasts of quarter-on-quarter (q-o-q) rates of real GDP growth and year-on-year (y-o-y) inflation rates of the consumer price index (CPI).[12] The quarterly forecasts are usually gathered in the first half of the last month of a quarter. Therefore, the forecasters can be expected to have information about the variable of interest in the current quarter, i.e. for the forecast (resp. nowcast) horizon $h = 0$. Given the y-o-y definition, and denoting the forecast horizon for the current quarter, i.e. for the nowcast with $h = 0$, we can expect that $h^* \geq 2$ for CPI inflation. This is because knowledge about past values of the price index enables the forecasters to mechanically produce forecasts which have lower MSPEs than the unconditional mean up to $h = 2$.[13] The countries under study are the US, the euro area, Japan, Germany, the UK, Italy, Canada, and France. Since, in each quarter, Consensus Economics also provides data for recent quarters, we can employ this real-time data for the evaluation of the forecasts. We use second vintages for both variables.

Considering forecasts for up to $h = 6$ quarters ahead, our evaluation sample mostly starts in the second quarter of 2001 and ends in the second quarter of 2018, yielding a sample size of $n = 69$. Only for the euro area, the sample starts in the second quarter of 2004, leading to $n = 57$. For the recursive mean serving as a benchmark forecast, the estimation begins with the $T = 20$ observations before the start of the evaluation sample.[14]

The empirical maximum forecast horizons $\widehat{h}^*$ determined by the tests are shown in Table 5. The sequential p-values of the tests giving rise to these values of $\widehat{h}^*$ are displayed in Figures 1 to 4. These figures also contain the ratios of the survey forecasts' MSPEs to the MSPEs of the respective benchmark forecasts.

[11] Results of the tests $dm_{0,h}$, $dm_{T,h}$, and the test of Giacomini and White (2006) for real GDP growth in the US can be found in the online appendix.

[12] For the UK, we use forecasts of the retail price index (RPI) because of their larger sample size.

[13] The year-on-year rate for $h = 2$ equals the sum of the quarter-on-quarter rates for $h = -1, 0, 1, 2$. Using the observed quarter-on-quarter rate for $h = -1$ and the unconditional mean as the forecast of the quarter-on-quarter rates for the latter three horizons yields an MSPE for the year-on-year rate forecast for $h = 2$ which is lower than the variance of the year-on-year rates by construction. If information on the current quarter is available, the maximum forecast horizon should thus be equal to or larger than 3.

[14] The recursive and the in-sample mean employ the same second-vintage realizations as used for the forecast evaluations.

[Table 5 about here.]

Notably, for GDP growth, $\widehat{h}^*$ is always smaller than the largest forecast horizon of $h = 6$. The tests for $\beta_h = 0$ and $\gamma_h = 0$ mostly yield results between $\widehat{h}^* = 1$ and $\widehat{h}^* = 3$ with a median result of about 2. For 5 out of 8 countries, both tests give identical results, while for the US, the euro area, and Japan, $\widehat{h}^*$ is 1 or 2 quarters larger when $\gamma_h = 0$ is used instead of $\beta_h = 0$. The findings of Section 5 suggest that this may be due to size distortions of the test for $\gamma_h = 0$ in the scenario with small noise.

For many countries, the tests for $\beta_h = 0.5$ and $\gamma_h = 0.5$ stop rejecting the null hypotheses about 1–2 quarters before their respective counterparts which test for equality to 0. $\widehat{h}^*$ mostly lies in the range 0–2. However, for Japan, the null hypotheses cannot even be rejected for the nowcast. The results of both tests coincide for 4 out of 8 countries. For 3 countries, the test for $\gamma_h = 0.5$ yields a value of $\widehat{h}^*$ which exceeds the value obtained with the test for $\beta_h = 0.5$ by 1 quarter, while for Italy, the opposite is observed.

To sum up, the survey forecasts for GDP often are not significantly more accurate than simple unconditional mean forecasts except at very short horizons. To be more precise, using the *no information* hypothesis, most real GDP growth forecasts turn out not to be informative for more than 1 quarter ahead. Yet, also at larger horizons, a linear transformation of the survey forecasts would often yield lower MSPEs than the unconditional means. In the latter sense, i.e. based on the *constant mean* hypothesis, the survey forecasts for GDP growth are informative for 2–3 quarters ahead in the majority of cases.

[Figure 1 about here.]

[Figure 2 about here.]

For CPI inflation, we find larger values of $\widehat{h}^*$ as expected due to the y-o-y definition. Concerning the tests for $\beta_h = 0$ and $\gamma_h = 0$, $\widehat{h}^*$ equals 3 or 4 in most cases. However, for Japan and Italy, the tests even reject for the last horizon, implying that $\widehat{h}^*$ at least equals 6. This result might be related to large changes of the value-added tax rate (VAT) which are commonly announced well in advance.[15] For 5 out of 8 countries, both tests again give identical results,

[15] For instance, the Japanese VAT rate increased from 5% to 8% in the second quarter of 2014. This pre-announced increase equals about 3 times the standard deviation of Japanese CPI inflation and, thus, leads to extremely large forecast errors of the benchmark forecasts, but not of the survey forecasts.

while the test for $\gamma_h = 0$ yields larger values of $\widehat{h}^*$ for 3 countries. For the US and France $\widehat{h}^*$ equals 4 instead of 3. For Canada, $\widehat{h}^*$ at least equals 6 if the test of $\gamma_h = 0$ employed, which is considerably larger than $\widehat{h}^* = 3$ as obtained with the test of $\beta_h = 0$. As Figures 3 and 4 show, Canada's ratios of the survey forecasts' MSPEs to the MSPEs of the respective benchmark forecasts are essentially flat and very close to 1 for $h \geq 4$. This suggests that small differences in the testing approaches can lead to relatively large differences concerning $\widehat{h}^*$.

When the tests for $\beta_h = 0.5$ and $\gamma_h = 0.5$ are employed, we again find values of $\widehat{h}^*$ that are often 1–2 quarters smaller than when using $\beta_h = 0$ and $\gamma_h = 0$. While $\widehat{h}^*$ continues to equal 3 or 4 in the majority of cases, values as small as 2 occur for the US, Germany and the UK, and $\widehat{h}^* \geq 6$ is only observed for Italy. Both tests give identical results for 6 out of 8 countries, while for the US and Germany, the test for $\gamma_h = 0.5$ yields $\widehat{h}^* = 3$ instead of $\widehat{h}^* = 2$ with the test for $\beta_h = 0.5$.

Thus, the survey forecasts for CPI inflation mostly outperform unconditional mean forecasts only at horizons where the y-o-y definition gives the survey forecasts an informational advantage. Only rarely do the tests find more accurate survey forecasts for $\widehat{h}^* > 3$. For several countries, however, employing linear transformations of the survey forecasts would again yield lower MSPEs than employing the unconditional means. Accordingly, based on the *constant mean* hypothesis, the survey forecasts for CPI inflation are usually found to be informative for 3–4 quarters ahead.

[Figure 3 about here.]

[Figure 4 about here.]

7 Conclusions

This paper develops a forecast evaluation framework for testing the null hypothesis that the forecast at some pre-specified horizon h is uninformative. We consider three different scenarios: In the first scenario the forecast is identical to some conditional mean whereas in the second scenario some noise is superimposed on the conditional mean. The third scenario relates to model-based forecasts where the parameters of the model are estimated in a recursive manner. For the first scenario a Diebold-Mariano type test statistic is proposed that performs very well

in our Monte Carlo experiments. For the empirically more realistic second and third scenario we adopt the encompassing principle that yields simple regression-based test statistics.

While all regression-based tests work reasonably well in the majority of cases considered, the tests using the recursive mean as a benchmark (tests based on γ_h) can suffer from non-negligible size distortions in certain situations. They also require more information than the tests based on the in-sample mean (tests based on β_h). Furthermore, we think that testing for a coefficient equal to zero is more appealing than testing for 0.5. First, the former test has more power and, second, it is invariant to linear transformations of the forecast and therefore, for instance, robust to forecast bias.

In the empirical analysis, we apply our tests to macroeconomic forecasts from the survey of Consensus Economics. Our results suggest that forecasts of macroeconomic key variables are hardly informative beyond 2–4 quarters ahead. Our results confirm earlier findings from the macroeconomic forecasting literature which were based on less rigorous approaches. The main contribution of our work is to provide statistical tests that allow the forecaster to assess the maximum forecast horizon of the forecast of interest.

It is worth mentioning that our testing approach (as any other empirical methodology) has two major limitations. First, the estimated maximum forecast horizon may be biased downwards whenever the power of the test is poor (e.g. for a small number of forecasts in the evaluation sample). Second, the estimated maximum forecast horizon depends on the approach that generates the forecasts. If the approach fails to exploit important information it may produce uninformative forecasts, while a richer forecasting procedure may result in informative forecasts. Accordingly, any qualification of the informative content is conditional on the forecasting approach.

References

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

Ang, A., G. Bekaert, and M. Wei (2007): "Do macro variables, asset markets, or surveys forecast inflation better?," *Journal of Monetary Economics*, 54(4), 1163–1212.

Breitung, J., and M. Knüppel (2018): "How far can we forecast? Statistical tests of the predictive content," Discussion Papers 07/2018, Deutsche Bundesbank.

Calhoun, G. (2016): "An asymptotically normal out-of-sample test based on mixed estimation windows," *Iowa State University, mimeo*.

Carlson, J. A., and M. J. Parkin (1975): "Inflation Expectations," *Economica*, 42(166), 123–138.

Cheng, M., N. Swanson, and C. Yao (2020): "Forecast Evaluation," in *Macroeconomic Forecasting in the Era of Big Data*, ed. by P. Fuleky, chap. 16, pp. 495–537. Springer.

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

Clark, T. E., and K. D. West (2007): "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138(1), 291 – 311, 50th Anniversary Econometric Institute.

Clements, M. P. (2019): *Macroeconomic survey expectations*, Palgrave texts in econometrics. Palgrave Macmillan, Cham, Switzerland.

Diebold, F. X., and L. Kilian (2001): "Measuring predictability: theory and macroeconomic applications," *Journal of Applied Econometrics*, 16(6), 657–669.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–63.

Elliott, G., and A. Timmermann (2016): *Economic Forecasting*. Princeton University Press, 1 edn.

Galbraith, J. W., and G. Tkacz (2007): "Forecast content and content horizons for some important macroeconomic time series," *Canadian Journal of Economics*, 40(3), 935–953.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74(6), 1545–1578.

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

Isiklar, G., and K. Lahiri (2007): "How far ahead can we forecast? Evidence from cross-country surveys," *International Journal of Forecasting*, 23(2), 167–187.

Knüppel, M. (2018): "Forecast-Error-Based Estimation of Forecast Uncertainty When the Horizon Is Increased," *International Journal of Forecasting*, 34(1), 105–116.

McCracken, M. W. (2019): "Tests of Conditional Predictive Ability: Some Simulation Evidence," Working Papers 2019-11, Federal Reserve Bank of St. Louis.

Mincer, J. A., and V. Zarnowitz (1969): "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, ed. by J. A. Mincer, chap. 1, pp. 1–46. NBER.

Nelson, C. R. (1976): "The Interpretation of R^2 in Autoregressive-Moving Average Time Series Models," *The American Statistician*, 30(4), 175–180.

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

Patton, A. J., and A. Timmermann (2012): "Forecast Rationality Tests Based on Multi-Horizon Bounds," *Journal of Business & Economic Statistics*, 30(1), 1–17.

Appendix A: Tests against a noninformative benchmark

In this appendix we analyse the tests invoking the recursive mean $\overline{Y}_t$ as uninformative benchmark. As noted in Remark 4, this hypothesis can be tested by running the regression

$$Y_{t+h} - \overline{Y}_t = \gamma_h \left(\widehat{Y}_{t+h|t} - \overline{Y}_t \right) + \nu_{t+h}, \tag{16}$$

where $\overline{Y}_t$ denotes the recursive mean of the expanding sample $\{Y_{-T+1}, \ldots, Y_t\}$. The LM version of the HAC test statistic is constructed as in (13), where

$$\begin{aligned} a_t &= \left[Y_{t+h} - \overline{Y}_t - 0.5(\widehat{Y}_{t+h|t} - \overline{Y}_t) \right] (\widehat{Y}_{t+h|t} - \overline{Y}_t) \quad \text{for } H_0 : \gamma_h = 0.5 \\ a_t &= \left(Y_{t+h} - \overline{Y}_t \right) (\widehat{Y}_{t+h|t} - \overline{Y}_t) \quad \text{for } H_0 : \gamma_h = 0. \end{aligned}$$

The following theorem presents the limiting null distribution of this test.

Theorem A.2 *Under Assumptions 1 – 2, $h > h^*$ and $\sigma_\eta^2 > 0$ the HAC t-statistics constructed as in (13) possess a limiting standard normal distribution as $n \to \infty$ and $T \to \infty$.*

Proof: See online appendix.

Finally we note that the test of hypothesis $\gamma_h = 0$ in (16) is equivalent to the adjusted MSPE test for nested forecast comparisons proposed by Clark and West (2007). Clark and McCracken (2001) showed that under the conditions of Theorem 3 (in particular $n/T \to 0$) the HAC t-statistic possesses a standard normal limiting null distribution, whereas the test is slightly conservative whenever n/T is substantial.

Appendix B: Proofs of the main results

Proof of Theorem 1:

(i) For the first statistic $dm_{0,h}$ we have

$$\begin{aligned} \delta_{0,t}^h &= (Y_{t+h} - \mu)^2 - (Y_{t+h} - \overline{Y}^h)^2 \\ &= u_{h,t}^2 - (u_{h,t} - \overline{u}_h)^2 \\ &= 2u_{h,t}\overline{u}_h - \overline{u}_h^2 \\ \sum_{t=1}^{n} \delta_{0,t}^h &= n\overline{u}_h^2 \end{aligned}$$

where $\overline{u}_h = n^{-1} \sum_{t=1}^{n} u_{h,t}$. This in turn yields

$$\frac{1}{\omega_h^2} \sum_{t=1}^{n} \delta_{0,t}^h \xrightarrow{d} \chi^2.$$

(ii) Under the null hypothesis we have for the statistic $dm_{T,h}$

$$\begin{aligned}
\frac{1}{\omega_h^2}\sum_{t=1}^{n}\delta_{T,n}^{h} &= \frac{1}{\omega_h^2}\sum_{t=1}^{n}\left\{(Y_{t+h}-\mu)^2-\left[Y_{t+h}-\mu-(\overline{Y}_t-\mu)\right]^2\right\} \\
&= -\frac{1}{\omega_h^2}\sum_{t=1}^{n}(\overline{Y}_t-\mu)^2+\frac{2}{\omega_h^2}\sum_{t=1}^{n}(Y_{t+h}-\mu)(\overline{Y}_t-\mu) \\
&= -\frac{1}{\omega_h^2}\sum_{t=1}^{n}(T+n)(\overline{Y}_t-\mu)^2\frac{1}{T+n}+\frac{2}{\omega_h^2}\sum_{t=1}^{n}\left[\sqrt{T+n}(\overline{Y}_t-\mu)\right]\left[\frac{1}{\sqrt{T+n}}(Y_{t+h}-\mu)\right] \\
&\Rightarrow -\int_{\pi}^{1}\frac{1}{a^2}W(a)^2da+2\int_{\pi}^{1}\frac{1}{a}W(a)dW(a)
\end{aligned}$$

where $\Rightarrow$ denotes weak convergence with respect to the associated probability measure, $\pi = T/(T+n)$, and

$$\begin{aligned}
\sqrt{T+n}\ (\overline{Y}_t-\mu) &= \frac{T+n}{T+t}\frac{\left(\sum_{s=-T-h+1}^{t-h}u_{h,s}\right)}{\sqrt{T+n}} \\
&= \frac{T+n}{T+t}\frac{\left(\sum_{s=-T+1}^{t-1}u_{h,s}\right)}{\sqrt{T+n}}+O_p[(T+n)^{-1/2}] \\
&\Rightarrow \frac{\omega_h}{a}W(a)
\end{aligned}$$

with $a=(T+t)/(T+n)$ and $W(a)$ is a standard Brownian motion.

Proof of Lemma 1:

Under the null hypothesis and Assumptions 1 and 2 we have

$$\mathbb{E}(\widehat{\overline{Y}}_h)=\mathbb{E}(\overline{Y}^h)-\mathbb{E}\left(n^{-1}\sum_{t=1}^{n}u_{h,t}\right)+\mathbb{E}\left(n^{-1}\sum_{t=1}^{n}\eta_t\right)=\mu$$

and the least-squares estimator of β_h in (11) is a consistent estimator of

$$\beta_h=\frac{\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\left(\widehat{Y}_{t+h|t}-\mu\right)(Y_{t+h}-\mu)\right]}{\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left(\widehat{Y}_{t+h|t}-\mu\right)^2}.$$

From

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{E}\left[(Y_{t+h} - \widehat{Y}_{t+h|t})^2 - (Y_{t+h} - \mu)^2\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left\{\left[(Y_{t+h} - \mu) - (\widehat{Y}_{t+h|t} - \mu)\right]^2 - (Y_{t+h} - \mu)^2\right\} \\
&= \sum_{t=1}^{n} \mathbb{E}\left[(\widehat{Y}_{t+h|t} - \mu)^2 - 2(Y_{t+h} - \mu)(\widehat{Y}_{t+h|t} - \mu)\right] \\
&= \mathbb{E}\left(\sum_{t=1}^{n}(\widehat{Y}_{t+h|t} - \mu)^2\right)\left[1 - 2\frac{\mathbb{E}\left(\sum_{t=1}^{n}(Y_{t+h} - \mu)(\widehat{Y}_{t+h|t} - \mu)\right)}{\mathbb{E}\left(\sum_{t=1}^{n}(\widehat{Y}_{t+h|t} - \mu)^2\right)}\right] \\
&= (1 - 2\beta_h)\sum_{t=1}^{n} \mathbb{E}\left[(\widehat{Y}_{t+h|t} - \mu)^2\right]
\end{aligned}$$

it follows that the null hypothesis (2) is equivalent to testing $\beta_h = 0.5$ in regression (11).

For testing the same hypothesis based on the recursive mean as the uninformative benchmark we define

$$\gamma_h = \frac{\sum_{t=1}^{n} \mathbb{E}\left[(Y_{t+h} - \overline{Y}_t)(\widehat{Y}_{t+h|t} - \overline{Y}_t)\right]}{\sum_{t=1}^{n} \mathbb{E}\left[(\widehat{Y}_{t+h|t} - \overline{Y}_t)^2\right]}.$$

Using

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{E}\left[(Y_{t+h} - \widehat{Y}_{t+h|t})^2 - (Y_{t+h} - \overline{Y}_t)^2\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left\{\left[(Y_{t+h} - \overline{Y}_t) - (\widehat{Y}_{t+h|t} - \overline{Y}_t)\right]^2 - (Y_{t+h} - \overline{Y}_t)^2\right] \\
&= \sum_{t=1}^{n} \mathbb{E}\left[(\widehat{Y}_{t+h|t} - \overline{Y}_t)^2 - 2(Y_{t+h} - \overline{Y}_t)(\widehat{Y}_{t+h|t} - \overline{Y}_t)\right] \\
&= (1 - 2\gamma_h)\sum_{t=1}^{n} \mathbb{E}\left[(\widehat{Y}_{t+h|t} - \overline{Y}_t)^2\right].
\end{aligned}$$

it follows that under Assumptions 1 – 2, the null hypothesis implies $\gamma_h = 0.5$.

Proof of Theorem 2:

Under the null hypothesis (2) and Assumptions 1 – 2 we have

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[(Y_{t+h}-\widehat{Y}_{t+h|t})^2-(Y_{t+h}-\mu)^2\right]\\
&=\frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n}(u_{h,t}-\eta_t)^2-(\mu_{h,t}-\mu+u_{h,t})^2\right]\\
&=\sigma_\eta^2-\sigma_\mu^2
\end{aligned}$$

where $\sigma_\eta^2=\mathbb{E}(n^{-1}\sum_{i=1}^{n}\eta_t^2)$ and $\sigma_\mu^2=\mathbb{E}[n^{-1}\sum_{i=1}^{n}(\mu_{h,t}-\mu)^2]$. It follows that the null hypothesis (2) implies $\sigma_\eta^2\geq\sigma_\mu^2$.

The test statistic for $\beta_h=0.5$ is constructed by using

$$\begin{aligned}
a_t&=(Y_{t+h}-\overline{Y}^h)(\widehat{Y}_{t+h|t}-\overline{\widehat{Y}}_h)-\frac{1}{2}(\widehat{Y}_{t+h|t}-\overline{\widehat{Y}}_h)^2\\
&=(\widetilde{\mu}_{h,t}+\widetilde{u}_{h,t})(\widetilde{\eta}_t+\widetilde{\mu}_{h,t})-\frac{1}{2}(\widetilde{\eta}_t+\widetilde{\mu}_{h,t})^2\\
&=\frac{1}{2}(\widetilde{\mu}_{h,t}^2-\widetilde{\eta}_t^2)+\widetilde{u}_{h,t}(\widetilde{\eta}_t+\widetilde{\mu}_{h,t}),
\end{aligned}$$

where a tilde above the symbol indicates a mean adjusted series, e.g. $\widetilde{\mu}_{h,t}=\mu_{h,t}-n^{-1}\sum_{s=1}^{n}\mu_{h,s}$. Under the hypothesis $\sigma_\eta^2=\sigma_\mu^2$ we have

$$\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\widetilde{\mu}_{h,t}^2-\widetilde{\eta}_t^2)\right)=0.$$

Assumption 1 and Assumption 2 (ii) imply that $u_{h,t}$ is uncorrelated with $\mu_{n,t}$ and η_t. Thus

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\widetilde{u}_{h,t}\widetilde{\mu}_{h,t}\right)&=\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{h,t}\mu_{h,t}\right)+\mathbb{E}\left(\sqrt{n}\;\overline{u}_h\overline{\mu}_h\right)\;=\;O_p(n^{-1/2})\\
\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\widetilde{u}_{h,t}\widetilde{\eta}_t\right)&=\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{h,t}\eta_t\right)+\mathbb{E}\left(\sqrt{n}\;\overline{u}_h\overline{\eta}\right)\;=\;O_p(n^{-1/2}),
\end{aligned}$$

where $\overline{u}_h=n^{-1}\sum_{t=1}^{n}u_{h,t}=O_p(n^{-1/2})$, $\overline{\mu}_h=n^{-1}\sum_{t=1}^{n}(\mu_{h,t}-\mu)=O_p(n^{-1/2})$, and $\overline{\eta}=n^{-1}\sum_{t=1}^{n}\eta_t=O_p(n^{-1/2})$. It follows that

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}a_t\right)=0.$$

Assumptions 1 and 2 ensure that the sample covariance $n^{-1}\sum_{t=j+1}^{n}a_ta_{t-j}$ converges in probability to its expectation for any finite j and provide the requirements for the Lindeberg-Feller central limit theorem. Therefore the test statistic constructed as in (13) possesses a standard normal limiting distribution.

For the null hypothesis $\beta_h = 0$ and Assumptions 1 and 2 we obtain

$$\begin{aligned}
&\sum_{t=1}^{n} a_t \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[\widehat{Y}_{t+h|t}(Y_{t+h} - \mu)\right] \\
&\qquad = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[(\eta_t + \mu_{h,t})(\mu_{h,t} - \mu + u_{h,t})\right] \\
&\qquad = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[\mu_{h,t}(\mu_{h,t} - \mu)\right] \;\; = \;\; 0
\end{aligned}$$

which implies $\mu_{h,t} = \mu$, that is, the *constant mean hypothesis* (3). Under the null hypothesis we therefore have

$$\sum_{t=1}^{n} (Y_{t+h} - \overline{Y}^h)(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h) = \sum_{t=1}^{n} \widetilde{u}_{h,t}\widetilde{\eta}_t$$

Using previous results for the hypothesis $\beta_h = 0.5$ it follows immediately that

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} a_t\right) = 0$$

and the sample covariances of a_t converge to their expectations. Accordingly, the test statistic has a standard normal limiting distribution.

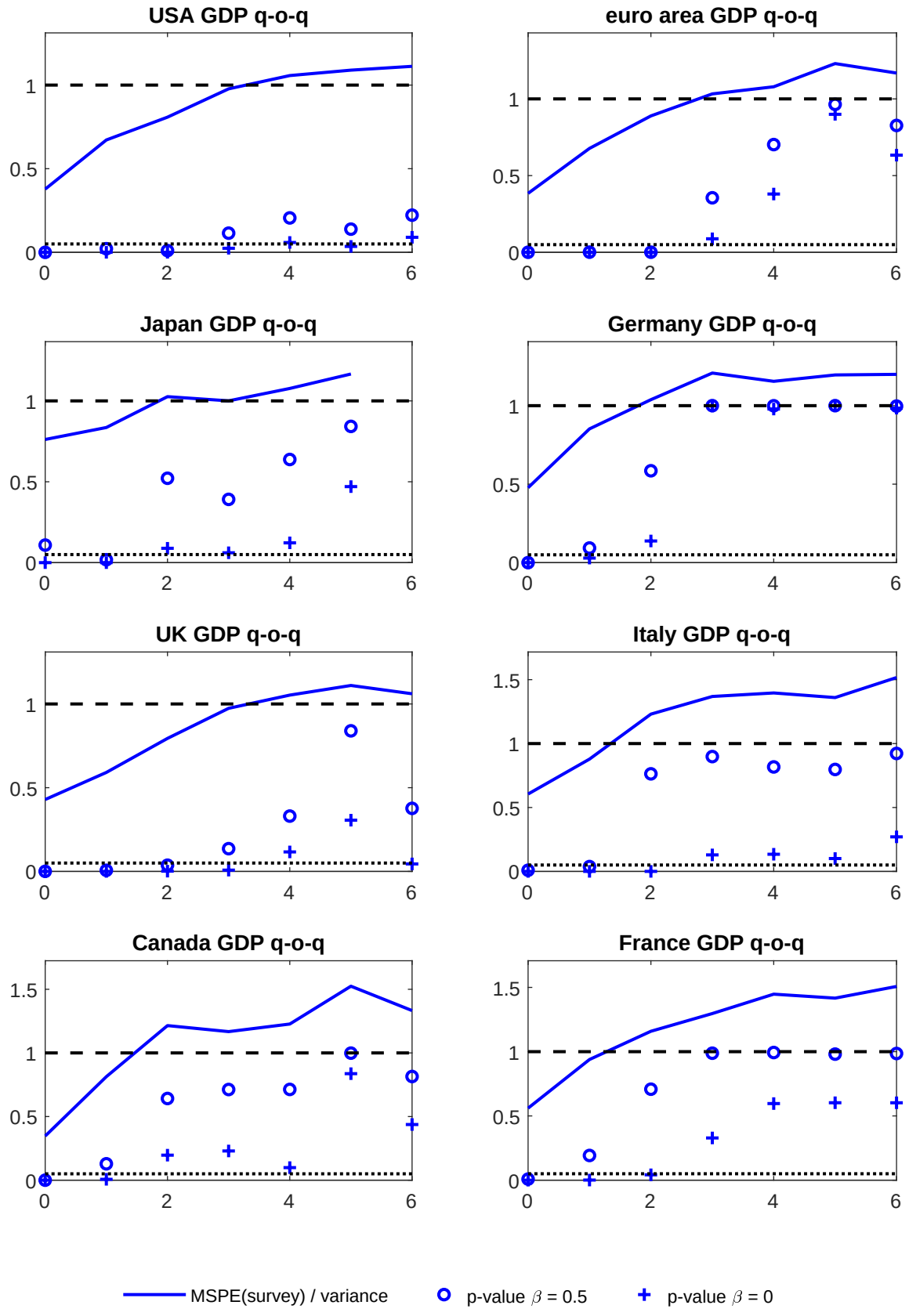

Figure 1: Test results for quarter-on-quarter growth rates of real GDP using the in-sample mean as the benchmark. The number on the x-axis denotes the forecast horizon in quarters with 0 being the nowcast. The dotted line is at 0.05, corresponding to the significance level of the tests. The dashed line is at 1. The solid line indicates the ratio of the Consensus forecasts' MSPE to the variance. All tests are one-sided. The respective values of $\widehat{h}^*$ are displayed in Table 5.

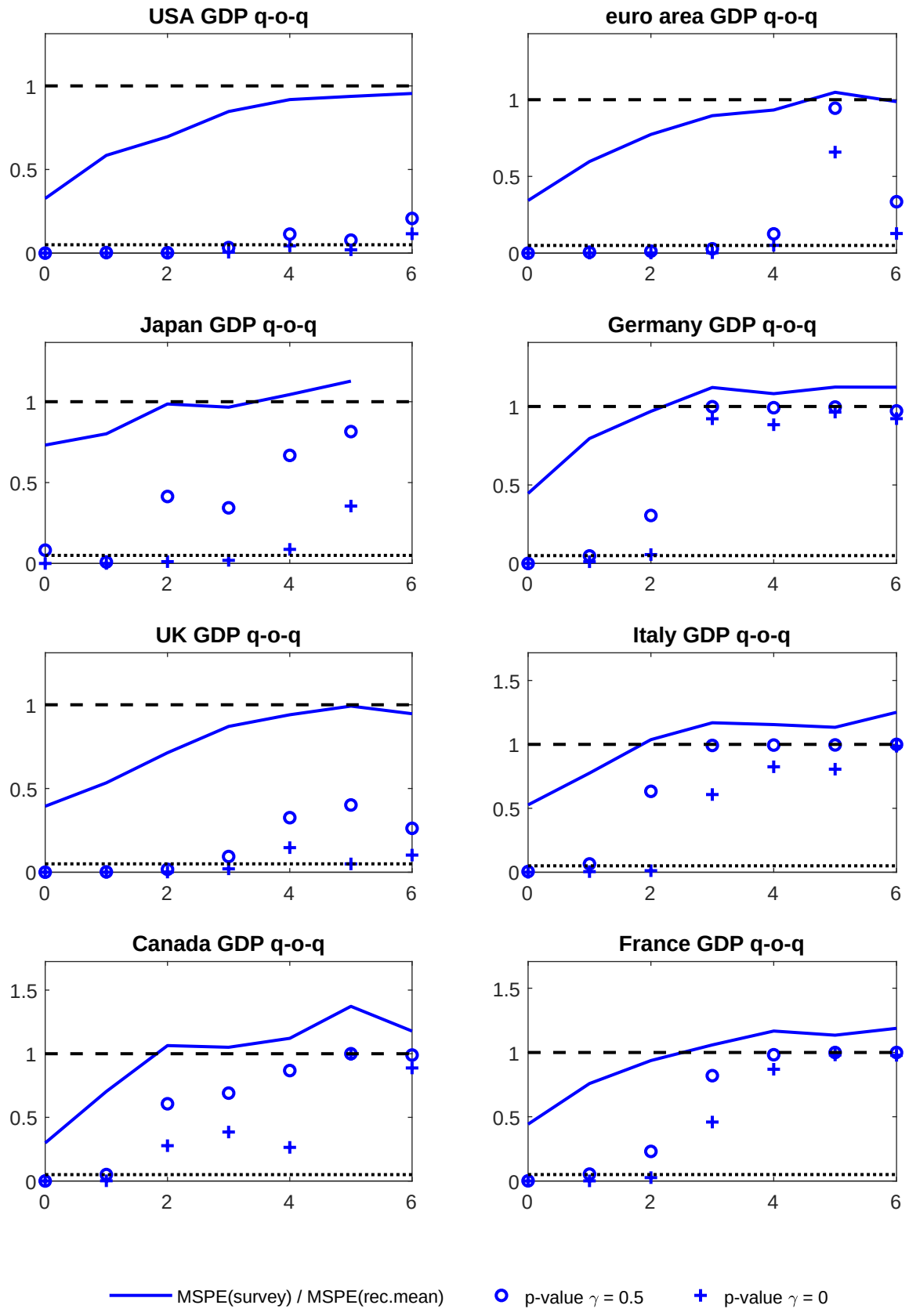

Figure 2: Test results for quarter-on-quarter growth rates of real GDP using the recursive mean with $T = 20$ initial observations as the benchmark. The solid line indicates the ratio of the Consensus forecasts' MSPE to the MSPE of the recursive mean forecasts. For further explanations, see Figure 1.

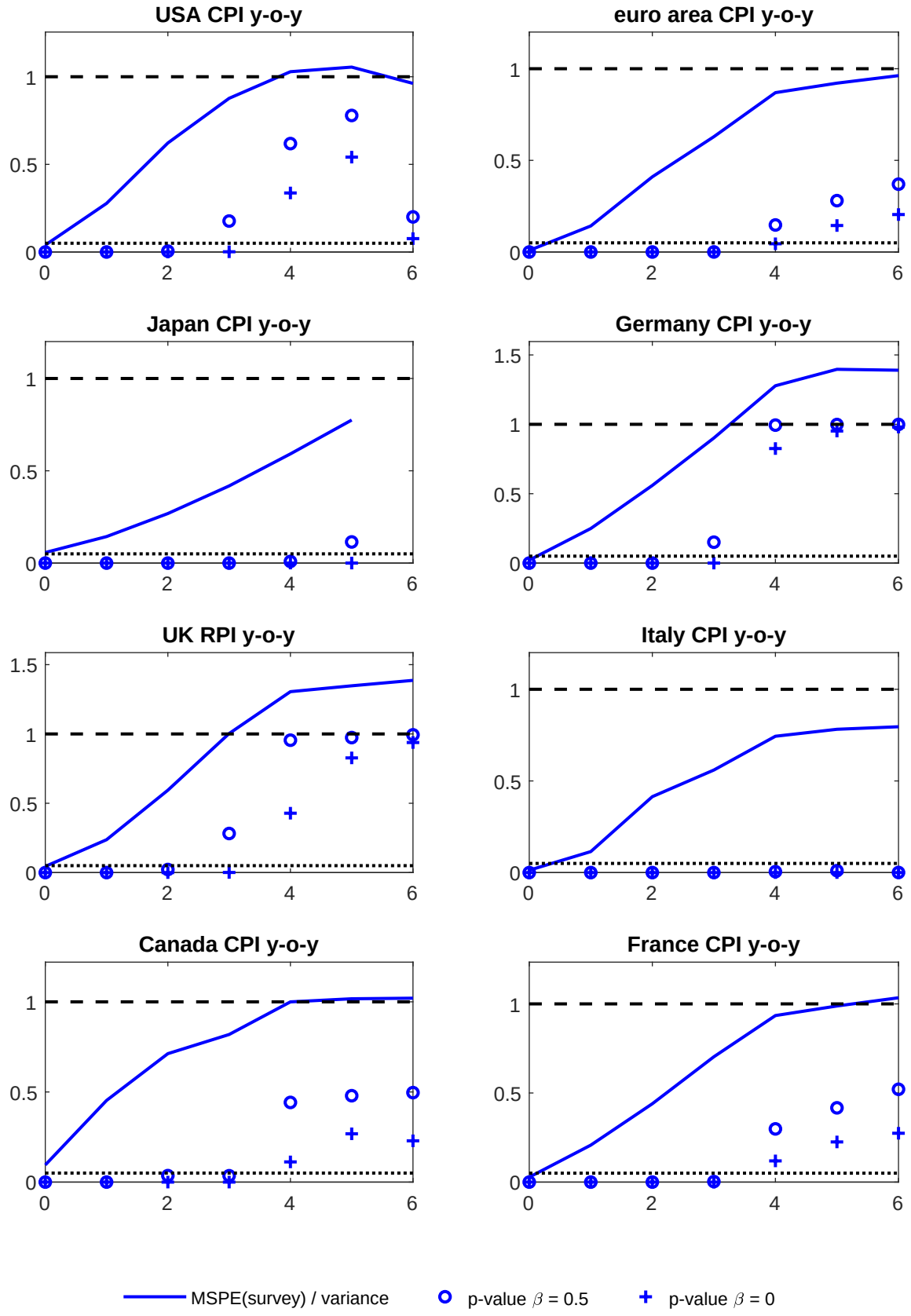

Figure 3: Test results for year-on-year growth rates of the CPI (the RPI in the case of the UK) based on the in-sample mean as the benchmark.. For further explanations, see Figure 1.

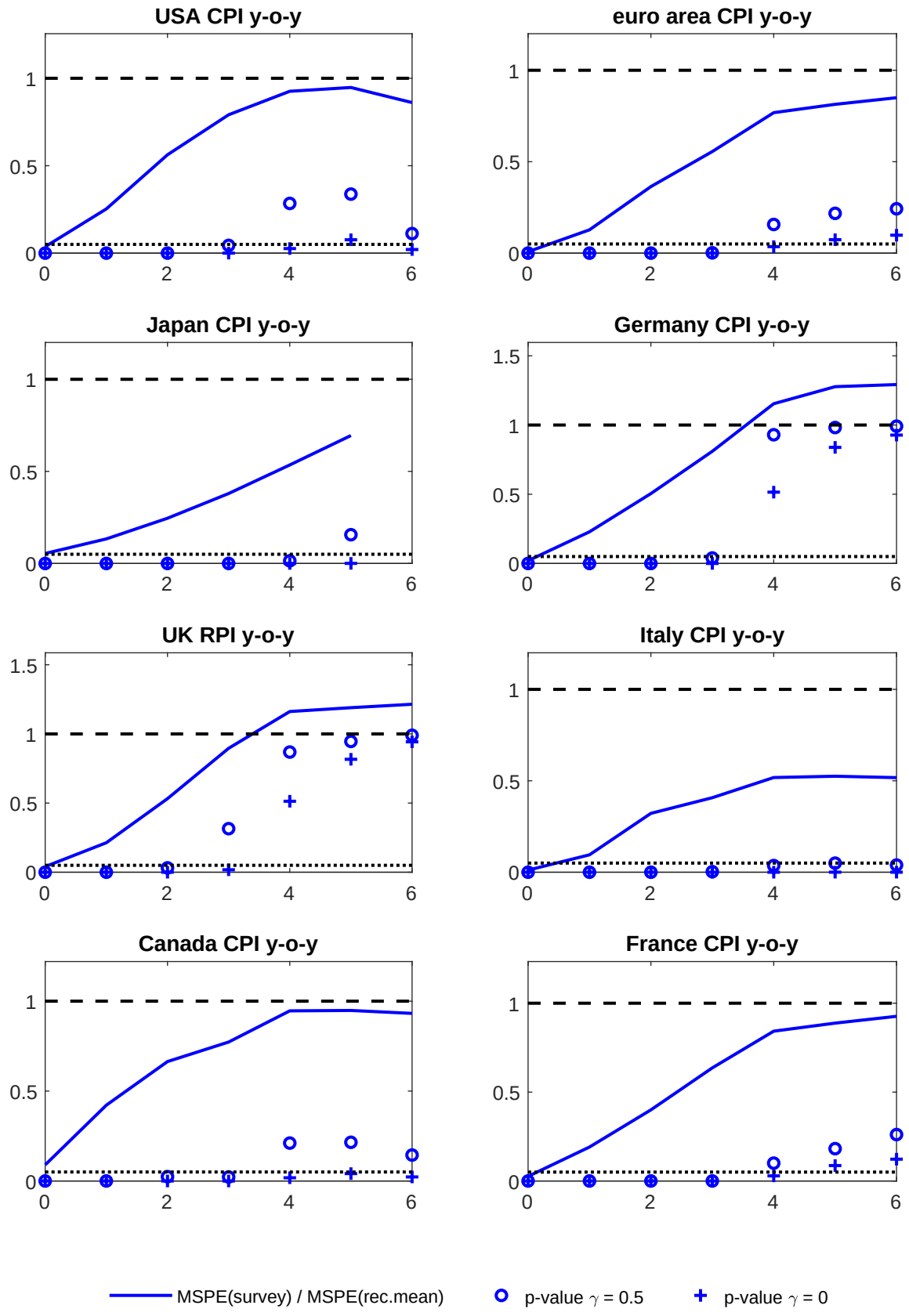

Figure 4: Test results for year-on-year growth rates of the CPI (the RPI in the case of the UK) using the recursive mean with $T = 20$ initial observations as the benchmark. For further explanations, see Figures 1 and 2.

Table 1: Actual size of tests of Theorems 1 (w/o noise) and 2 (w/ noise)

n	25	50	100	250	500	25	50	100	250	500	25	50	100	250	500
								$b = 0$							
								$\sigma_\eta = 0$							
								dm_0							
						0.05	0.05	0.05	0.05	0.05					
								dm_T							
$T = 50$						0.06	0.05	0.05	0.05	0.05					
$T = 100$						0.06	0.05	0.05	0.05	0.05					
$T = 250$						0.06	0.05	0.05	0.05	0.05					
$T = 500$						0.05	0.05	0.05	0.05	0.05					
$T = 1000$						0.07	0.05	0.05	0.05	0.05					
	$\sigma_\eta = 0.001$					$\sigma_\eta = 0.01$					$\sigma_\eta = 0.1$				
			dm_0					dm_0					dm_0		
	0.05	0.05	0.05	0.06	0.07	0.10	0.12	0.13	0.16	0.18	0.22	0.22	0.20	0.15	0.10
			dm_T					dm_T					dm_T		
$T = 50$	0.06	0.05	0.05	0.06	0.05	0.06	0.06	0.05	0.06	0.05	0.09	0.08	0.08	0.07	0.05
$T = 100$	0.06	0.05	0.05	0.06	0.05	0.06	0.06	0.05	0.06	0.05	0.11	0.10	0.09	0.08	0.05
$T = 250$	0.06	0.06	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.05	0.15	0.13	0.11	0.08	0.06
$T = 500$	0.05	0.06	0.05	0.06	0.05	0.05	0.05	0.05	0.06	0.06	0.19	0.17	0.14	0.10	0.06
$T = 1000$	0.06	0.05	0.05	0.05	0.05	0.07	0.05	0.06	0.06	0.05	0.26	0.21	0.17	0.12	0.07
			$\beta = 0$					$\beta = 0$					$\beta = 0$		
	0.06	0.06	0.06	0.05	0.05	0.07	0.06	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.05
			$\gamma = 0$					$\gamma = 0$					$\gamma = 0$		
$T = 50$	0.20	0.24	0.30	0.44	0.58	0.20	0.23	0.29	0.42	0.54	0.13	0.15	0.16	0.19	0.20
$T = 100$	0.17	0.19	0.23	0.34	0.44	0.15	0.18	0.22	0.30	0.39	0.10	0.11	0.12	0.15	0.15
$T = 250$	0.13	0.15	0.17	0.24	0.30	0.11	0.13	0.15	0.21	0.26	0.08	0.08	0.08	0.10	0.10
$T = 500$	0.11	0.12	0.14	0.19	0.23	0.10	0.10	0.11	0.15	0.19	0.07	0.07	0.07	0.08	0.09
$T = 1000$	0.10	0.10	0.12	0.15	0.18	0.08	0.08	0.09	0.12	0.15	0.06	0.06	0.06	0.07	0.07
								$b = \sigma_\eta$							
	$\sigma_\eta = 0.001$					$\sigma_\eta = 0.01$					$\sigma_\eta = 0.1$				
			$\beta = 0.5$					$\beta = 0.5$					$\beta = 0.5$		
	0.06	0.06	0.05	0.05	0.05	0.06	0.06	0.05	0.06	0.05	0.06	0.06	0.05	0.05	0.05
			$\gamma = 0.5$					$\gamma = 0.5$					$\gamma = 0.5$		
$T = 50$	0.14	0.14	0.15	0.18	0.20	0.13	0.14	0.15	0.16	0.17	0.08	0.08	0.08	0.09	0.08
$T = 100$	0.12	0.13	0.14	0.16	0.18	0.11	0.11	0.13	0.14	0.15	0.07	0.07	0.08	0.08	0.08
$T = 250$	0.11	0.11	0.12	0.14	0.16	0.09	0.09	0.10	0.11	0.12	0.07	0.06	0.06	0.07	0.06
$T = 500$	0.09	0.10	0.11	0.12	0.14	0.08	0.08	0.08	0.10	0.10	0.06	0.06	0.06	0.06	0.06
$T = 1000$	0.09	0.09	0.10	0.11	0.13	0.08	0.07	0.07	0.08	0.09	0.06	0.06	0.06	0.06	0.05

Note: Test results for data-generating process $Y_t = b\,X_{t-1} + \epsilon_t$ with $X_t, \epsilon_t \overset{iid}{\sim} \mathcal{N}(0,1)$ and 10,000 simulations. Forecasts are given by $\widehat{Y}_{t+1|t} = b\,X_t + \eta_t$ with $\eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2)$. dm_0 and the tests for β are based on in-sample means of Y_t. Benchmark forecasts for tests based on dm_T and γ are estimation-sample means of Y_t using a recursive estimation scheme starting with T observations. Significance level is $\alpha = 0.05$. Tests are based on OLS standard errors. All tests are one-sided. The subscript h is suppressed in $dm_{0,h}$, $dm_{T,h}$, β_h and γ_h, because $h = 1$ for all tests.

Table 2: Power of tests of Theorems 1 (w/ noise) and 2 (w/o noise) for noisy forecasts

n	25	50	100	250	500	25	50	100	250	500	25	50	100	250	500
								$b = 0.2$							
			$\sigma_\eta = 0.001$					$\sigma_\eta = 0.01$					$\sigma_\eta = 0.1$		
			dm_0												
	0.52	0.63	0.77	0.91	0.98										
			dm_T					dm_T							
$T = 50$	0.37	0.47	0.63	0.86	0.97	0.34	0.51	0.74	0.97	1.00					
$T = 100$	0.43	0.52	0.66	0.87	0.97	0.31	0.48	0.70	0.96	1.00					
$T = 250$	0.51	0.59	0.71	0.88	0.97	0.29	0.45	0.67	0.95	1.00					
$T = 500$	0.55	0.64	0.75	0.90	0.97	0.28	0.44	0.65	0.94	1.00					
$T = 1000$	0.61	0.67	0.78	0.91	0.98	0.28	0.43	0.65	0.94	1.00					
			$\beta = 0$					$\beta = 0$					$\beta = 0$		
	0.28	0.41	0.64	0.93	1.00	0.28	0.42	0.64	0.93	1.00	0.25	0.36	0.56	0.88	0.99
			$\beta = 0.5$					$\beta = 0.5$					$\beta = 0.5$		
	0.14	0.18	0.27	0.47	0.73	0.15	0.18	0.27	0.48	0.72	0.11	0.13	0.17	0.28	0.44
													$\gamma = 0.5$		
$T = 50$											0.13	0.15	0.21	0.34	0.50
$T = 100$											0.12	0.14	0.20	0.32	0.49
$T = 250$											0.11	0.14	0.18	0.31	0.47
$T = 500$											0.11	0.13	0.18	0.29	0.46
$T = 1000$											0.11	0.13	0.18	0.29	0.45

Test results for data-generating process $Y_t = b\,X_{t-1} + \epsilon_t$ with $X_t, \epsilon_t \overset{iid}{\sim} \mathcal{N}(0,1)$ and 10,000 simulations. Results are only displayed for tests with actual size < 0.10 for all n, T considered in Table 1. Forecasts are given by $\widehat{Y}_{t+1|t} = b\,X_t + \eta_t$ with $\eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2)$. dm_0 and β indicates tests based on in-sample means of Y_t. Benchmark forecasts for tests based on dm_T and γ are sample means of Y_t using a recursive estimation scheme starting with T observations. Significance level is $\alpha = 0.05$. Tests are based on OLS standard errors. All tests are one-sided. The subscript h is suppressed in $dm_{0,h}$, $dm_{T,h}$, β_h and γ_h, because $h = 1$ for all tests.

Table 3: Actual size and power of tests of Theorem 3 (model predictions)

n	25	50	100	250	500	25	50	100	250	500
	$b = 0$									
	$\beta = 0$					$\gamma = 0$				
$T = 50$	0.02	0.02	0.02	0.01	0.01	0.04	0.03	0.03	0.02	0.02
$T = 100$	0.03	0.02	0.02	0.01	0.01	0.04	0.03	0.03	0.03	0.02
$T = 250$	0.03	0.02	0.02	0.02	0.01	0.04	0.04	0.03	0.03	0.03
$T = 500$	0.04	0.03	0.02	0.02	0.01	0.04	0.04	0.03	0.03	0.03
$T = 1000$	0.04	0.03	0.03	0.02	0.02	0.05	0.04	0.04	0.03	0.03
	$b = 0.2$									
	$\beta = 0$					$\gamma = 0$				
$T = 50$	0.16	0.24	0.41	0.78	0.97	0.20	0.31	0.50	0.86	0.99
$T = 100$	0.20	0.30	0.48	0.83	0.98	0.23	0.35	0.56	0.89	0.99
$T = 250$	0.26	0.36	0.57	0.89	0.99	0.27	0.38	0.61	0.91	0.99
$T = 500$	0.27	0.39	0.60	0.91	1.00	0.27	0.40	0.62	0.92	1.00
$T = 1000$	0.28	0.41	0.62	0.93	1.00	0.28	0.42	0.63	0.93	1.00

Note: Test results for data-generating process $Y_t = b\,X_{t-1} + \epsilon_t$ with $X_t, \epsilon_t \overset{iid}{\sim} \mathcal{N}(0,1)$ and 10,000 simulations. Forecasts are given by $\widehat{Y}_{t+1|t} = \widehat{a} + \widehat{b}\,X_t$ with $\widehat{a}, \widehat{b}$ resulting from OLS regressions of Y_t on a constant and X_{t-1}. Benchmark forecasts for tests based on γ are estimation-sample means of Y_t. Both forecasts use a recursive estimation scheme starting with T observations. Tests based on β are considered in Theorem 2. Significance level is $\alpha = 0.05$. Tests are based on OLS standard errors. All tests are one-sided. The subscript h is suppressed in β_h and γ_h, because $h = 1$ for all tests.

Table 4: Giacomini-White tests for unconditional forecast comparisons

n		25	50	100	250	500	25	50	100	250	500
		$\widetilde{dm}_B = 0$									
		$b = b_{GW}$					$b = 0.2$				
	b_{GW}										
$B = 50$	0.1458	0.08	0.04	0.02	0.02	0.02	0.11	0.08	0.06	0.08	0.14
$B = 100$	0.1012	0.07	0.05	0.03	0.02	0.02	0.14	0.14	0.14	0.21	0.39
$B = 250$	0.0655	0.08	0.06	0.04	0.02	0.02	0.17	0.18	0.23	0.36	0.63
$B = 500$	0.0458	0.08	0.06	0.05	0.03	0.02	0.19	0.20	0.26	0.44	0.70
$B = 1000$	0.0323	0.09	0.06	0.05	0.04	0.03	0.19	0.21	0.28	0.46	0.72

Note: Test results for data-generating process $Y_t = b\,X_{t-1} + \epsilon_t$ with $X_t, \epsilon_t \overset{iid}{\sim} \mathcal{N}(0,1)$ and 10,000 simulations. Forecasts are given by $\widehat{Y}_{t+1|t} = \widehat{a} + \widehat{b}\,X_t$ resulting from OLS regressions of Y_t on a constant and X_{t-1}. Benchmark forecasts are estimation-sample means of Y_t. Both forecasts use a rolling estimation scheme with B observations. Significance level is $\alpha = 0.05$. Tests are based on Newey and West (1987) standard errors with truncation lags chosen according to Andrews (1991). The values of b_{GW} used for the size simulations are calibrated to yield identical mean-squared prediction errors (MSPE) of $\widehat{Y}_{t+1|t}$ and the rolling sample means. The MSPEs are 1.0417 ($B = 50$), 1.0205 ($B = 100$), 1.0080 ($B = 250$), 1.0040 ($B = 500$) and 1.0020 ($B = 1000$), respectively. These values of the MSPEs are based on 1 billion simulations and have standard errors of about 0.00005, respectively. The test is one-sided. The subscript 1 is suppressed in $\widetilde{dm}_{B,1}$.

Table 5: Empirical maximum forecast horizons $\widehat{h}^*$ in quarters for forecasts of growth and inflation

	US	EA	JP	DE	UK	IT	CA	FR	median
					GDP q-o-q				
$\beta_h = 0$	3	2	1	1	3	2	1	2	2
$\gamma_h = 0$	5	3	3	1	3	2	1	2	2.5
$\beta_h = 0.5$	2	2	−1	0	2	1	0	0	0.5
$\gamma_h = 0.5$	3	3	−1	1	2	0	0	0	0.5
					CPI y-o-y				
$\beta_h = 0$	3	4	6	3	3	6	3	3	3
$\gamma_h = 0$	4	4	6	3	3	6	6	4	4
$\beta_h = 0.5$	2	3	4	2	2	6	3	3	3
$\gamma_h = 0.5$	3	3	4	3	2	6	3	3	3

Note: 'GDP q-o-q' denotes quarter-on-quarter growth rates of real GDP, 'CPI y-o-y' year-on-year growth rates of the consumer price index except for the UK, where the retail price index is employed. The abbreviations used for the countries are 'US' for the United States, 'EA' for the euro area, 'JP' for Japan, 'DE' for Germany, 'UK' for the United Kingdom, 'IT' for Italy, 'CA' for Canada, and 'FR' for France. Forecast errors cover the sample 2004q2 to 2018q2 for the euro area and 2001q2 to 2018q2 for all other countries. Forecasts and real-time observations (second vintage) are taken from Consensus Economics. The benchmark mean forecasts for the tests of $\gamma = 0$ and $\gamma = 0.5$ are based on recursive estimations starting with $T = 20$ observations. $\widehat{h}^*$ is set to -1 if the null hypothesis cannot be rejected for the nowcast. A value of 6 implies that the maximum forecast horizon equals *at least* 6 quarters, because 6 quarters is the largest forecast horizon under study.