Norges Bank

12 | 2025

Working Paper

Cross-Check of Economic Forecasts

Norges Bank Research

Authors:
Frida Bowe
Eleonora Granziera
Pål B. Ulvedal

Keywords
Forecasting, Forecast evaluation, DSGE modelling, Bayesian VAR

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles på e-post: servicesenter@norges-bank.no

Fra 1999 og senere er publikasjonene tilgjengelige på http://www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail: servicesenter@norges-bank.no

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-8143 (online)

ISBN 978-82-8379-373-4 (online)

Cross-Check of Economic Forecasts*

Frida Bowe† Eleonora Granziera‡ Pål B. Ulvedal §

September 29, 2025

Abstract

Policymakers cross-check their projections for multiple variables and forecast horizons with experts' forecasts or satellite models. This paper proposes a set of *quantitative metrics* that can be used to summarize the overall discrepancy between two forecasting models *jointly* across variables and forecasting horizons. The methodologies can handle situations where only the point forecast is available as well as where the full predictive densities are known. It also allows to take into account the policymaker loss function, by assigning different weights to variables or horizons. We illustrate the usefulness of our measures when comparing the forecasts from the Survey of Professional Forecasters, the Tealbook, a medium scale Bayesian VAR, and a medium scale Dynamic Stochastic General Equilibrium (DSGE) model for the U.S. data. We find that the forecasts substantially depart ahead of and during recessions, resulting in our measures spiking.

Keywords: Forecasting, Forecast Evaluation, DSGE modelling, Bayesian VAR

JEL classification: C32, C52, C53, E17

*This working paper should not be reported as representing the views of Norges Bank. The views expressed are those of the author and do not necessarily reflect those of Norges Bank. We would like to thank Sona Benecka (discussant), and conference participant at the 28th International Conference Computing in Economics and Finance, the Danmarks Nationalbank, Deutsche Bundesbank and Norges Bank "The Return of Inflation", the 29th International Conference on Macroeconomic Analysis and International Finance for helpful comments and suggestions.

†Norges Bank; email: frida.bowe@norges-bank.no

‡Norges Bank; email: eleonora.granziera@norges-bank.no

§Nord University; email: pal.b.ulvedal@nord.no

1 Introduction

Forecasting future economic conditions is essential for the conduct of monetary policy. Policymakers often rely on different models to produce forecasts for several different variables and horizons of interest. Projections can form the basis for monetary policy decisions and affect expectations of financial markets, experts and the general public (Granziera et al. (2025)). The information content of the forecasts may vary, from point forecasts, interval forecasts, to full predictive densities.

Central banks may have either one main model or a suite of models to produce staff forecasts, run counterfactual exercises, and conduct scenario analysis. When a central bank operates with a main model, it is often selected based on several criteria beyond just forecast accuracy, such as its capacity to interpret new information and maintain consistency across various forecasts. Consequently, other models may have better forecasting properties for certain variables. Alternatively, staff forecasts can be obtained by combining several models using statistical techniques and expert judgment. In both cases, it is important to cross-check in real time the staff projections with other sets of forecasts, either obtained from other internal models, produced by other institutions, or collected through surveys. This should reduce the chances of making large and systematic forecasting errors, which might undermine central bank credibility.

For example, the report on economic conditions and monetary policy prepared for the Federal Open Market Committee (FOMC) by the Board of Governors staff (Tealbook) compares the Tealbook forecasts for real activity, labor market conditions, interest rates, and three measures of inflation with Blue Chips and Survey of Professional Forecasters (SPF) forecasts from nowcast up to two years ahead, as shown in figure (1). Similarly, Norges Bank's projections for the Norwegian economy are produced with its main Dynamic Stochastic General Equilibrium (DSGE) model and cross-checked with the forecasts from a Bayesian Vector Autoregression (BVAR) model, as well as several other models. Before finalizing the forecasts during the projection process, several iterations between the empirical cross-check models and the main policy model are conducted.[1]

In this context, a significant challenge for the staff is to summarize for the policymaker the discrepancies across all variables and horizons and to determine whether these discrepancies are substantial enough to warrant further investigation. Forecasts are often made for a wide range of variables and horizons, making it difficult to easily and succinctly identify where significant discrepancies exist between model forecasts.

This paper suggests methodologies for comparing the forecasts from a benchmark model

[1] See Norges Bank's Monetary Policy Handbook for more information (Norges Bank, 2022, p. 71).

with those from an alternative model or with external forecasts in real time, during the projection round when the forecasts are made. Rather than assessing each variable and forecasting horizons individually, the objective is to look at them *jointly*, therefore summarizing the information in a single metric or statistic. In other words, our suggested measures help the policymaker to interpret the overall discrepancy between the different forecasting models across variables and horizons. The methodologies are developed to take into account several possible scenarios, depending on whether only the point forecasts, or interval or full predictive densities are available.

A first simple measure of discrepancy is the Euclidian norm, which computes the distance across variables and forecasting horizons. It is straightforward to compute and requires only knowledge of the two sets of point forecasts. The norm is reminiscent of the Root Mean Squared Forecast Error (RMSFE), where the squared difference across forecasts is replaced by the squared forecast error. However, differently from measures of accuracy, which can be computed only ex-post as they require the knowledge of the realized value of the target variables, the norm can be computed in real time, because its only inputs are the forecasts. A drawback of the norm is that it disregards the correlation of the forecasts among variables and forecast horizons: for example, the norm could be very high because of a large gap in the forecasts observed for only one variable at the nowcast horizon, which might carry over to other forecasting horizons if the variable is very persistent. Moreover, it is hard to determine the threshold that characterizes alarming large values. To overcome this second shortcoming, we suggest to examine the norm relative to its history and provide some simple rules to define the threshold. Our recommendation is to use an average of the norm computed over a rolling window of past values.

An alternative approach to compare point forecasts, that overcomes the drawbacks of the norm, is the Wald test, which treats the one set of point forecasts as the null hypothesis, and the second set of forecasts as data. The advantage of the Wald test over the norm is that it offers a clear cut-off value depending on the significance level set by the researcher: if the test rejects the null, then the researcher can conclude that the forecasts are statistically significantly different from each other. A challenge of this measure is that it requires the estimation of the covariance matrix of the forecasts, therefore, the researcher needs a long evaluation sample to precisely estimate it. The difficulty in estimating the covariance matrix increases with the number of variables and the forecasting horizons included in the cross-check. The use of shrinkage or factor-based covariance estimators mitigates the issue.

Last, we suggest a measure to compare point forecasts to interval forecasts. Based on the work of Christoffersen (1998), we consider the correct coverage. This measure can be implemented whenever either the full posterior distribution or interval forecasts for one set

of forecasts are available. The main idea is to assume that the posterior distribution from one set of forecasts is the "true" forecast density. For a single given variable and forecast horizon, the researcher computes an indicator function that takes the value of one if the forecast for the same variable and horizon is within the credible set or forecast interval of the "true" model, and zero otherwise. Then, the coverage tells us the proportion of variables and forecasting horizons for which the point forecast from the alternative model is included in the credible set of the true model. As for the Wald test, the correct coverage requires the researcher to specify a level of significance, which will be used to construct the forecast intervals.

The metrics can be easily modified to take into account the policymaker loss function, by assigning different weights to variables and horizons. For example, a policymaker might be more concerned if the discrepancies involve forecasts of inflation and output rather than house prices, for example, as these may lead to diverging policy recommendation. Alternatively, given the lags in the transmission of monetary policy, central bankers might prefer to focus on discrepancies occurring at long forecast horizons.

We illustrate the usefulness of our measures by comparing forecasts from the Survey of Professional Forecasters, the Tealbook, a medium scale Bayesian VAR, and a medium scale DSGE (Dynamic Stochastic General Equilibrium) model using U.S. data spanning the last forty years. We compute our measures and report time series of the norm and coverage from 1981Q3 and of the Wald-test from 1995Q1, as for this measure we need to set aside several vintages of forecasts to estimate the covariance matrix. We observe that the time series for all our metrics exhibit significant spikes during and ahead of recessions, indicating substantial deviations in the forecasts during these periods. These findings hold true when comparing model forecasts, models with expert judgment and different experts' based forecasts. Notably, the highest spikes for the norm and for the Wald test are observed during the pandemic, when, for all variables and horizons we observe significant deviations. Although the Wald test and the coverage show fewer spikes than the norm, they still exhibit clear spikes during recessions. This suggests that economic turmoil, such as the Great recession, leads to greater divergence in model forecasts. Similarly, expert judgment, as in the Tealbook and SPF, may have a harder time forecasting future economic developments during these periods.

Note that the cross-checks suggested in this paper are not just academic exercises, but are highly policy relevant, given the ongoing shift among central banks from reliance on a single "core" model to the systematic use of suites of models. In the pandemic and post-pandemic environment—characterized by elevated inflation, and large shocks—institutions have increasingly complemented their baseline projections with satellite models, and survey-based

forecasts. Therefore, real-time cross-check metrics are essential safeguards that help ensure robust policy deliberation, and reduce the risk of groupthink. Our results show that tracking dispersion across models and surveys serves as an early-warning indicator of downturns and illustrate how cross-check metrics organize conflicting evidence into actionable guidance so that policymakers navigate uncertainty in real time, maintain coherence across forecasts and narratives, and make better-informed policy choices.

Our paper contributes to the vast literature on forecast evaluation. On the theoretical side, most of the papers on this topic focus on absolute performance, i.e. bias or efficiency (e.g. Nordhaus (1987), Holden and Peel (1990)), or relative performance, by comparing forecast accuracy (e.g. Diebold and Mariano (1995), Giacomini and White (2006)). Only a handful of papers compare the forecasts from two models jointly across multiple horizons (Capistrán (2006), Patton and Timmermann (2012), Quaedvlieg (2021)). Most of the literature, instead, is concerned with comparison across several models for one single variable and one single forecast horizon (White (2000), Hansen (2005), Hansen (2011), Clark and McCracken (2012), Granziera et al. (2014)). Regarding empirical findings, one complementary paper to ours is Granziera and Sekhposyan (2019), which shows that the relative accuracy of different forecasting models for inflation and industrial production changes over time and large divergence in accuracy is associated with economic recessions. Though related, note that the evidence in Granziera and Sekhposyan (2019) is based on forecast errors, which can be computed only ex-post, while our measures can provide a real time warning to policymakers.

The focus of this paper is to determine whether two sets of forecasts are broadly similar, or very different from each other, rather than to assess their forecast accuracy. We therefore suggest measures which can summarize the discrepancy in the forecasts *at a glance*, without having to analyze separately each single graph, as those shown in Figure (1). We note that a large divergence might be a sign of trouble and warrant further investigation. Also, the set of models to include in the cross-check can be chosen based on their past forecasting performance. The metrics are therefore meant to complement, rather than substitute, model analysis focused on accuracy.

Evidence collected from survey-based measure of expectations shows that disagreement in the cross-section of agents regarding future realizations of inflation correlates with rises in mean inflation and predicts periods of elevated uncertainty or economic distress (Brandao-Marques et al. (2024), Tsiaplias (2020)). We provide complementary evidence showing that discrepancy between different models predicts future turmoil.

The rest of the paper is organized as follows: section 2 presents the methodologies. Section 3 describes the data, models and forecasts included in the empirical application.

Section 4 discusses the results and suggests an algorithm to carry out the cross-check. Section 5 concludes.

2 Methodology

In this section we suggest formal and systematic ways to cross-check two sets of forecasts over several variables and multiple forecast horizons. We first show how to compare point forecasts. Those can be computed as point predictions from a time series model, or they could be model-free forecasts, such as mean forecasts from survey data, or published projections from central banks or other institutions. Then, we move to compare point with interval forecasts, which can be easily obtained from predictive densities of time series models, such as BVARs, once the desired level of significance is selected. Alternatively, they could be obtained from fan charts of published projections, for example by the Bank of England or Norges Bank, or by the cross-sectional distribution of survey-based expectations.[2]

2.1 Comparing Point Forecasts

Denote by $\overline{y}^b_{t+h,k|t}$, and $\overline{y}^d_{t+h,k|t}$ the point forecasts for variable k, made at time t for forecasting horizon h. The superscripts b and d denote two sets of forecasts. Those could be obtained from survey data, time series models or structural models.

Norm. The simplest approach in summarizing the "disagreement" between two sets of point forecasts at time t consists in computing the distance across variables and horizons through the Euclidian norm:

$$norm_t = d\left(\overline{Y}^b_{t+H,K|t}, \overline{Y}^d_{t+H,K|t}\right) = \left\|\overline{Y}^b_{t+H,K|t}, \overline{Y}^d_{t+H,K|t}\right\| = \left[\sum_{h=1}^{H}\sum_{k=1}^{K}|\overline{y}^b_{t+h,k|t} - \overline{y}^d_{t+h,k|t}|^2\right]^{1/2} \tag{1}$$

where $\overline{Y}^i_{t+H,K|t}$ is a vector that stacks the point forecast for variables $k = 1, .., K$, horizons $h = 1, ..., H$ and model $i = \{b, d\}$ made at time t. Then, the norm is just the squared root of the sum of the squared differences between the forecasts of the two models across variables and across forecast horizons. As such, it weighs positive and negative deviations equally and penalizes large deviations proportionally more than small deviations. This measure is straightforward to compute and requires only knowledge of the point forecasts $\overline{Y}^b_{t+H,K|t}$

[2]Note, however, that the cross-sectional dispersion of survey-based measures of expectations represent disagreement among respondents rather than uncertainty around the point prediction.

and $\overline{Y}^{d}_{t+H,K|t}$. It can only take positive values and it is not bounded above. A higher value implies larger discrepancy between the forecasts. It disregards the correlation between different variables and horizons. The former will likely be high whenever variables covering similar concepts are included in the comparison, e.g. PCE and CPI inflation; the latter will be higher for variables that are highly persistent, e.g. interest rates. The variables included in the norm might have different units of account (levels vs growth rates), or display substantial differences in their volatility. Therefore, the deviations $\overline{y}^{b}_{t+h,k|t} - \overline{y}^{d}_{t+h,k|t}$ in (1) are standardized by the real-time standard deviation of the corresponding variable k computed on realized data up to time $t-1$, $\sigma_{y_{k,t-1}}$, before taking the sum across variables:

$$norm_t = \left[\sum_{h=1}^{H} \sum_{k=1}^{K} \frac{|\overline{y}^{b}_{t+h,k|t} - \overline{y}^{d}_{t+h,k|t}|^2}{\sigma^2_{y_{k,t-1}}} \right]^{1/2} \tag{2}$$

The Euclidian norm in (1) is reminiscent of the Root Mean Squared Forecast Error (RMSFE), where the squared difference across forecasts is replaced by the squared forecast error. However, differently from measures of accuracy, which can be computed only ex-post, at $t+H$, as they require the knowledge of the realized value of the target variables, the norm can be computed in real time, at t, because its only input are the forecasts.

While the norm has the advantage of being straightforward to compute, its main challenge is that it can take any value above or equal to zero, and it is hard to determine what characterizes an alarming large value. One therefore has to examine the norm relative to its history.

In this simple version, the norm weights each variable and forecast horizon equally. Therefore, it would assign the same importance to deviations regarding variables included in the monetary authority's mandate, e.g. inflation and output, relative to deviations about auxiliary variables, e.g. residential investment. The measure in equation (1) could be easily modified to assign different weights to each variable and forecast horizon:

$$d_{\omega}\left(\overline{Y}^{b}_{t+H,K|t}, \overline{Y}^{d}_{t+H,K|t}\right) = \left[\sum_{h=1}^{H} \sum_{k=1}^{K} \omega_{h,k} |\overline{y}^{b}_{t+h,k|t} - \overline{y}^{d}_{t+h,k|t}|^2 \right]^{1/2} \tag{3}$$

The weights could be assigned through a statistical criterion, such as the inverse of the historical RMSFE, or reflect policy makers preferences, e.g. shorter term forecasts might get more weight than longer term forecasts, or key variables such as output and inflation, might be assigned a larger weight than, e.g., government spending.

Wald Test. An alternative approach treats one set of forecasts, say "b", as the null

hypothesis, and the second set of forecasts, say "d", as data. Denote the stacked point forecasts from model/survey b and model/survey d respectively as:

$$vec\left(\overline{Y}^{b}_{t+1|t},..,\overline{Y}^{b}_{t+H|t}\right)=\overline{Y}^{b}_{t+H,K|t}=\beta_0 \tag{4}$$

$$vec\left(\overline{Y}^{d}_{t+1|t},..,\overline{Y}^{d}_{t+H|t}\right)=\overline{Y}^{d}_{t+H,K|t}=\hat{\beta}_t \tag{5}$$

Therefore, β_0 denotes the forecast vintage t from model "b", while $\hat{\beta}_t$ denotes the same vintage t but from model/survey "d". Then, we want to test whether the point forecasts from model d are equal to the ones from b:

$$H_0:\hat{\beta}_t=\beta_0 \tag{6}$$

Assume that we have available a sample of forecast vintages with origins at $t=R+1,...R+P$. If furthermore we assume $\sqrt{P}\left(\hat{\beta}_t-\beta_0\right)\xrightarrow{d}N\left(0,V\right)$ then a natural approach to test jointly the multiple hypotheses on the parameters $\hat{\beta}_t$ for the null in (6) is to use the Wald test. The Wald test statistic is given by:

$$\lambda_W=\left(\hat{\beta}_t-\beta_0\right)'\hat{V}^{-1}\left(\hat{\beta}_t-\beta_0\right) \tag{7}$$

Under the null in (6) the asymptotic distribution of the Wald statistic in (7) is chi-squared with HK degrees of freedom:

$$\lambda_W\xrightarrow{d}\chi^2_{(HK)} \tag{8}$$

Therefore, the Wald-test can be thought of as a modification of the forecast accuracy evaluation tests proposed in Diebold and Mariano (1995), Giacomini and White (2006) and Capistrán (2006), where the forecast errors are replaced by the deviations between the forecasts. Once obtained the test statistic (7), one can compute the associated p-value p_{λ_W}. Either the test-statistic λ_W or the p-value p_{λ_W} can be used as measures of discrepancy. The merit of using the p-value is that by construction, the measure will be bounded between zero and one, where smaller numbers will mean stronger evidence against the null, i.e. the two sets of forecasts are further apart. The advantage of the Wald test over the norm is that it offers a clear cut-off value depending on the significance level set by the researcher: if the test rejects the null, then the researcher can conclude that the forecasts are statistically significantly different from each other. Moreover, the inclusion of the variance co-variance matrix means that the test takes into account the correlation across variables and horizons,

unlike the norm. Forecasts with higher variances, i.e. longer horizons forecasts, or forecasts of more volatile variables, such as residential investment, receive less weight in the Wald test statistic.

One challenge with this approach is that the dimension of the matrix $\hat{V}$, the variance covariance matrix of the forecasts, is potentially large relative to the number of observations in the out-of-sample, P, and therefore might not be precisely estimated. Also, the researchers need to set aside a sub-sample of observations to estimate the matrix $\hat{V}$, while the norm can be computed just with only one vintage of forecasts. As in the case of the norm, the reason behind rejection of the null (i.e. which forecast horizons/variables show the largest discrepancy) has to be further investigated. A difficulty in shedding light on the reason behind the rejection is due to the fact that the difference in the forecasts are weighted by the variance-covariance matrix $\hat{V}$.

The policymaker loss function can be incorporated by replacing the matrix V in (7) with the weighting matrix $V_W = V + V_P$ where V_P represents the policymaker preferences. For example, higher values in V_P could be assigned to variables/horizons that are less of interest to the policymaker.

2.2 Comparing Interval to Point Forecasts

Coverage. This measure can be implemented whenever either the full posterior distribution or interval forecasts for one set of forecasts are available. The main idea is to assume that the posterior distribution from the set of forecasts b is the "true" forecast density. For a single given variable and forecast horizon, the indicator function will take the value one if the point forecast from model d for the same variable and horizon lies within the α credible set or forecast interval of model b, and zero otherwise. Then, the coverage tells us the proportion of variables and forecasting horizons for which the point forecast from model d is included in the credible set of model b.

This approach is based on the Christoffersen (1998)'s LR test of the correct coverage (α) which can be implemented as follows:

$$I^{\alpha}_{t,h,k} = \mathbb{1}\{y^{d}_{t+h,k|t} \in FI^{\alpha,b}_{t,h,k}\} \tag{9}$$

where, as before, $y^{d}_{t+h,k|t}$ is the point forecast from the set of forecasts d for variable k and horizon h and $FI^{\alpha,b}_{t,h,k}$ the highest posterior density interval/forecast interval of level $1-\alpha$

for model b. A suggested indicator is:

$$I_t^{\alpha} = \frac{1}{KH} \sum_{k=1}^{K} \sum_{h=1}^{H} I_{t,h,k}^{\alpha} \tag{10}$$

that is the percentage of variables and forecasting horizons for which the point forecasts of model d are included in the confidence intervals of model b. Similarly to the norm, this approach disregards the covariance across forecast horizons and variables. Also, by construction, the measure is bounded between zero and one. Note that this measure does not capture the magnitude by which one forecast "misses" the forecast interval of the other.

The indicator can be easily modified to take into account the policy maker preferences by adding a set of weights to (10):

$$I_t^{\alpha} = \sum_{k=1}^{K} \sum_{h=1}^{H} w_{h,k} I_{t,h,k}^{\alpha} \tag{11}$$

2.3 Discussion of Cross-Check Measures

In this subsections we discuss the main advantages and caveats associated with each of the three measures proposed above, all summarized in Table (1) where we also list their closest forecast accuracy evaluation analogues.

The norm is the most straightforward measure to compute, as it requires the smallest amount of information: only one vintage of forecasts, only point forecasts and no nuisance parameter. A first downside is that it does not take into account correlations among forecasts across variables and horizons. Therefore, the norm could spike because of a large discrepancy in the nowcast of one persistent variable, which determines large deviations for the subsequent forecast horizons. Similarly, large deviations among variables that cover similar economic concepts could cause a surge in the norm. This issue could be addressed by carefully selecting the variables included in the comparison, or by assigning different weights to variables or forecast horizons. Second, to become operational, the norm requires the researcher to set a threshold that identifies alarming values. Moreover, like other measures based on squared deviations, it might be sensitive to outliers.

The Wald p-value overcomes the first two shortcomings of the norm but it requires to choose the significance level. Moreover, it relies on the assumption of asymptotic gaussian distribution of the forecast differentials. A further challenge is the potentially burdensome estimation of the variance covariance matrix, which requires a sample of forecasts larger than the number of variables times the number of forecasting horizons, KH. While this issue can be mitigated by the use of shrinkage or factor-based covariance estimators, comparing two

large models might be problematic as the length of the sample necessary to carry out the estimation increases with the number of variables and forecasting horizons considered.

Finally, like the norm, the coverage has the advantage that it can be computed with only one vintage of forecasts. However, it requires the knowledge of the forecast interval for at least one model/set of forecasts. In addition, this measure relies on the correct specification of the forecast densities and it is sensitive to the choice of the confidence level, which determines the width of the intervals.

3 Empirical Application

In this section we describe the survey-based forecasts as well as the time series models considered in this study.

3.1 Models and Forecasts

SPF Forecasts. We use the median forecast from the SPF for a selection of variables.[3] All the data is available at a quarterly frequency. The variables included in the comparison are real GDP ($RGDP$), the implicit GDP deflator ($PGDP$), real consumption ($RCON$), real residential fixed investment ($RRESINV$), real nonresidential fixed investment ($RNRESIN$), the civilian unemployment rate ($UNEMP$), and the three-month treasury bill rate ($TBILL$). All variables are included from the first publication of forecasts. Consequently, we include $RRGDP$ and $PGDP$ from 1975, while $RCON$, $RRESINV$ and $NNRESIN$ and the $TBILL$ from 1981.

Tealbook. For the Tealbook, former Greenbook, we collect the forecasts for the macroeconomic variables that overlap with those in the SPF database.[4] Therefore, we include real ($RGDP$), the implicit GDP deflator ($PGDP$), residential fixed investment ($RRESINV$) and the civilian unemployment rate ($UNEMP$). We include forecasts from 1981Q3 until the latest vintage available, i.e. 2019Q4. We select the Tealbook forecast that matches the publication of the SPF forecasts most closely, typically the forecast produced towards the second month of the quarter, when available. If this is missing, we select the forecast produced on the third month of the quarter. In this respect, we are giving an informational advantage to the Tealbook forecasts.

[3]For more information on the data see Federal Reserve Bank of Philadelphia (2024a).
[4]For more information on the data see Federal Reserve Bank of Philadelphia (2024b).

BVAR. We estimate a BVAR model recursively on real time data, and produce forecasts for each iteration. The model is medium scale and includes seven variables.

The model can be expressed in the following form:

$$Y_t = C + A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + u_t \tag{12}$$

where Y_t is the vector of n endogenous variables and u_t is the vector of residuals. For estimation we use Bayesian techniques, and the priors are set as in (Giannone et al., 2015). We include four lags, as customary for quarterly data. The following seven variables are included in the medium-scale BVAR model:

$$Y_t = \begin{bmatrix} \Delta log(RGDP)_t \\ \Delta log(PGDP)_t \\ \Delta log(RCONS)_t \\ \Delta log(RRESINV)_t \\ \Delta log(RNRESIN)_t \\ \Delta log(EMP)_t \\ r_t/4 \end{bmatrix}$$

where Y_t is the vector of endogenous variables, $RGDP$ is real GDP, $PGDP$ is the implicit GDP deflator, $RCONS$ is real consumption, $RRESINV$ is real residential fixed investment, $RNRESIN$ is real nonresidential fixed investment, EMP is nonfarm payroll employment, and r_t is the 3-month T-bill rate.[5] The specific variables are chosen based on their compatibility with the Survey of Professional Forecasters (SPF). To handle the volatile period during the Covid pandemic, we implement the methods suggested by Lenza and Primiceri (2022). This consists in explicitly modeling the change in shock volatility in 2020Q1. Specifically, they suggest to weight observations inversely proportionally to their innovation variance. By adopting a "pandemic-adjusted" prior in the BVAR, we increase the model's responsiveness to severe shocks without compromising the systematic relations among the variables. The model is estimated on quarterly data for the US from 1965Q1 to 2023Q1 with an expanding window approach. The first forecast origin for both the BVAR and DSGE model is 1975Q1.

DSGE. Similarly, we estimate the Smets-Wouters model (Smets and Wouters, 2007) recursively on real time data. The model is estimated using Bayesian methods on quarterly data for the US from 1957:Q1 to 2023:Q1. The observable variables include the same seven variables as used by Smets and Wouters (2007):

[5] We download the data from the Philadelphia Fed website https://www.philadelphiafed.org/surveys-and-data/data-files

$$Y_t = \begin{bmatrix} \Delta log(RGDP)_t \\ \Delta log(PGDP)_t \\ \Delta log(RCONS)_t \\ \Delta log(RINV)_t \\ \Delta log(RWAGE)_t \\ log(HOURS)_t \\ r_t/4 \end{bmatrix}$$

where Y_t is the vector of observable variables, $RGDP$ is real GDP, $PGDP$ is the implicit GDP deflator, $RCONS$ is real consumption, $RINV$ is real investment, $RWAGE$ is real wages, $HOURS$ is hours worked, and r is the 3-month T-bill rate.[6] The observable variables are connected to the model variables through a measurement equation of the following form:

$$Y_t = Hv_t \tag{13}$$

The point and interval forecasts from the BVAR and DSGE models can be easily obtained. The general econometric framework is the following. Consider the posterior predictive distribution of a Bayesian VAR model or a DSGE model:

$$p\left(Y^i_{T+1:T+H}|Y_{1:T}\right) = \int p\left(Y^i_{T+1:T+H}|\theta, Y_{1:T}\right) p\left(\theta|Y_{1:T}\right) d\theta \tag{14}$$

where Y is a $K \times 1$ vector of endogenous variables, θ is a $L \times 1$ vector that collects the parameters, $p\left(\theta|Y_{1:T}\right)$ is the posterior distribution and $p\left(Y^b_{T+1:T+H}|\theta, Y_{1:T}\right)$ the likelihood function. The superscript i denotes objects computed from the BVAR model, then $i = b$ or the DSGE model, then $i = d$, while the subscript $t_1 : t_2$ indicate sequences from t_1 to t_2, e.g., $Y_{1:T}$ is shorthand for $Y_1, ..., Y_T$.

Point and interval forecasts can be computed by drawing $KH \times 1$-dimensional vectors $\{Y^{i,(s)}_{T+1:T+H}\}$ from the posterior, where the superscript s is an index denoting a single draw. Then, the point forecast is the average of the draws:

$$E\left(Y^b_{T+h|T}\right) = \int_{Y^b_{T+h}} Y^b_{T+h} p\left(Y^b_{T+h}|Y_{1:T}\right) dy^b_{T+h} \approx \frac{1}{nsim} \sum_{s=1}^{nsim} Y^{b,(s)}_{T+h} = \overline{Y}^b_{T+h|T} \tag{15}$$

[6]The data is downloaded from the FRED database, see Federal Reserve Bank of St. Louis (2024)

4 Results

4.1 Forecast Accuracy

Before presenting the different measures, we examine the accuracy of the models and of the survey-based forecasts. Table (2) shows the relative RMSFE as well as the significance for the Diebold and Mariano (1995) forecas accuracy test for the out of sample forecasts of the BVAR, the DSGE model and the Tealbook compared to the SPF forecasts, from one to four quarters ahead. A number greater than one suggests that the SPF forecasts are more accurate than the competing forecasts. The comparison between the SPF and time series models is carried over the sample 1981Q3-2023Q1, while for the cross-checks with the Tealbook we use forecasts up to 2019Q4, as these forecasts are released to the public with a five year delay.

The SPF outperforms the BVAR and the DSGE model for every variable and horizon, though the gains generally fall as the horizon increases. Professional forecasters are substantially more accurate in forecasting the Treasury Bill rate, due to their ability to interpret the forward guidance of the Federal Reserve regarding the policy rate during the zero lower bound episodes. The lower RMSFEs of the SPF forecasts compared to our BVAR and DSGE forecasts are not unexpected, given that the superior real-time forecasting ability of the SPF forecasts with respect to time-series models has been documented in several studies (Faust and Wright (2013), Crushore and Stark (2019)).

Table (2) confirms the superiority of the Tealbook forecasts compared to the SPF forecasts found in previous studies (Sims (2002)). However, gains are small, and, with few exceptions, not statistically significantly different from zero. This finding is consistent with recent evidence of a declining information effect of the Fed compared to the private sector (Hoesch et al. (2023)). Note also that for most quarters two vintages of Tealbook projections are available. In those instances, we select the latest vintage, therefore giving some informational advantage to the Fed staff.

These results are insightful, as central banks may have a core model, either time series or structural, that ensures that the forecasts are consistent and can be explained, and allows one to undertake scenario analysis. While forecast accuracy of the projections might not be the only goal of the core model, policymakers do not want to make large forecast errors, in order to establish and maintain credibility. Therefore, it is important to compare central bank projections with those of an accurate benchmark in real time. Which alternative model should be the benchmark for the cross-check? Performing a forecast evaluation of the core model relative to alternative benchmarks helps to identify the best model or external set of forecasts to use for the cross-check exercises described below.

We consider three types of comparisons for our cross-checks exercises: the first includes only experts' forecasts, which incorporate judgment and anticipated future effects of current macro or sectoral developments. The second compares to the experts' forecasts those from a pure time series model without judgment, based on past correlations between macroeconomic variables. Finally, the third one involves two time series models, where one imposes cross-equation restrictions.

4.2 Cross-Check Measures

Norm. First, we show the results for the norm. This simple measure rises in the presence of substantial discrepancies among variables, drawing attention to the most extreme values. Note that, because the variables included in the norm have different units of measure, as suggested in section (2.1) we divide them by the real-time variance of the corresponding variable before summing the squared differentials in (1), where the variance is computed on an expanding window.[7] This ensures that large deviations are not mechanically due to high volatility of certain series. Next, we need to select the cutoff value that should serve as a warning sign to trigger a cross-check of the model. Our suggestion is to compare the norm to its past values. The threshold should change over time, to explicitly handle structural breaks and regime shifts and to shield practitioners from arbitrary or outdated benchmarks. One simple approach that fit these criteria is to evaluate whether the current norm lies above its real time mean computed over a rolling window, obtained using observations up to the previous quarter. We suggest a ten year rolling window mean.[8]

Figure (2), (5) and (6) show the time series of the norm computed for different forecast vintages over time together with a ten year rolling window mean. We distinguish among three different cross-checks: SPFs versus Tealbook, SPF versus BVAR and BVAR versus DGSE. For example, the norm for the first quarter of 2015 in Figure (2) shows the overall distance between the SPF and the Tealbook forecasts made in 2015Q1 for all the variables considered, from nowcasting up to four quarters ahead. From a visual inspection of the three figures, two observations stand out: first, the series show substantial variation over time; second, the measure clearly surges in the quarters leading up to or coinciding with a recession, therefore signaling turmoil ahead. This second finding echoes with the broader literature that documents increases in disagreement among forecasts of economic agents measured in expectations surveys. We now look more in detail at each bimodel cross-check. In the case

[7] For instance, for forecasts made at time t for forecast horizon from nowcast to four quarters ahead, we divide the squared differentials in the forecasts for real output by the variance of *realized* real output computed from 1947Q1 up to time $t-1$.

[8] For the first 40 observations in the sample, we compute the threshold as an expanding window mean up to the previous quarter.

of the SPF vs Tealbook comparison, we note that the norm rises above its mean ahead of recessions and the second highest peak is registered during the Great Financial Crisis, with warning signs starting in 2006Q2.[9] The cross-check between SPF and the BVAR confirms a large divergence in forecasts during recessions. Moreover, the addition of the pandemic sample highlights the relevance of this event to the economy, compared to previous recessions. The norm spikes also at the very end of the sample, with the recent surge in inflation. These findings resemble very closely those obtained when computing the norm for the medium scale BVAR and the DSGE model.[10] Then, a first takeaway of the empirical findings is that the norm is an early warning signal of downturns.

Why is it useful to compute the norm rather than looking at multiple graphs plotting forecast for each variables separately, like those in figure (1)? We answer this question by looking at an example from the comparison between the SPF and Tealbook forecasts. To illustrate the usefulness of our proposed measure, we report in panel (b) of figure (3) the SPF and Tealbook forecasts made in 2007Q3 the quarter before the start of the recession, when the norm peaked abruptly. Looking at the individual variables and horizons, we note that the Tealbook were less optimistic than the SPF: the Fed staff was anticipating lower economic activity and higher unemployment, with the gap for these two variables growing for the more distant future. Both were expecting a drop in residential investment, but the Fed staff was projecting a larger fall. Consistent with a stronger economic outlook, SPF were also expecting higher inflation. Despite these differences, it is not clear whether these two sets of forecasts diverge overall. For example, the forecasts for unemployment differ by 0.15 at most, the ones for PGDP depart initially but they converge four quarters out, conversely, the ones for output almost coincide at the nowcast, but diverge at longer horizons, the ones for residential investment commove, though with a wedge. Therefore, it would be hard from this figure to draw conclusions regarding the discrepancy among the two sets of forecasts, while the norm gives a distinct warning signal, when compared with its time series. To further illustrate our point, we show in panel (a) of figure (3) the forecasts made in 2005Q3, where the norm is at one of its lowest levels. In that quarter, professional forecasters were less optimistic than the Federal Reserve about the growth prospects of the economy and more

[9] Because of the publication lag of the Tealbook forecasts, unfortunately, we cannot extend the comparison to include the latest recession.

[10] Here the norm is computed only over four variables: real output, GDP deflator, real consumption and the tBill. These are all the variables that enter both models. Other variables cover similar concepts, but are not directly comparable. For example, for labor market indicators, the BVAR includes employment, while the DSGE looks at wages and hours. For private investment, the BVAR includes real residential investment and real non-residential investment, while the DSGE uses real total investment. Using the same exact variables in the models would lead to uninteresting comparisons, as the forecasts would be extremely similar, absent any judgement.

worried about inflation, consistent with the effects of negative supply shocks, but the visual inspection does not allow to understand whether the overall divergence across variables and horizons is larger in 2005Q3 or 2007Q3. The differences in inflation are more apparent in 2005Q3, while the divergence in real output and residential investments are larger in 2007Q3. The norm instead reveals at a glance a larger value in 2007Q3.

What should a policy maker do when the norm surges above the threshold? We recommend plotting the time series of the standardized squared differentials for each variable summed across horizons to investigate the source of divergence, as we do in Figures (4), (7) and (8). This simple exercise uncovers some interesting findings. The first one is that different variables can trigger the spike in the norm during the recessionary episodes. In the 1981-1982 recession the professional forecasters and the Federal Reserve were disagreeing about the future realizations of all variables. In the early 1990s' recession instead, the increase in the norm was due to divergence in output forecasts. During the Great Financial Crisis, the spike in the norm was driven by the gap in the forecasts of real residential investment and, to a larger extent, the unemployment rate. This finding seems to contrast with the results in figure (3), from which one could conclude that the jump in the norm in 2007Q3 is caused only by differences in the forecasts for real residential investment. This is due to the fact that in figure (4) the squared differentials are standardized by the real time standard deviation of the variable. Therefore, even if the square difference in the residential investment forecasts between SPF and Tealbook were larger than those in unemployment, because of the larger volatility of investment, the standardized differentials were smaller. This shows that plotting the norm and the standardized squared differentials is a useful exercise to understand the relative magnitude of the differences in the forecasts, both across time and across variables.

The second insight is that the variables determining the surge in the norm might vary according to the models considered in the comparison. An inspection of the squared differentials for each variable in the BVAR vs DSGE cross-check, shown in figure (8) highlights that divergence for the tBill forecasts were driving the spike in the norm in the early 80s and the Great Recession of 2008, while real consumption was behind the spike in the early 90s and in 2001. Differences in the forecasts for real output and real consumption contributed the most to the surge in 2020, though all variables show large differentials during the recession triggered by the pandemic. Similar findings hold for the DSGE vs BVAR comparison. However, in figures (5) and (6) we also observe a spike at the very end of the sample, in 2022Q3. In the first case, this is due to a divergence in the forecast of the GDP deflator, which the BVAR, incorrectly, projected to be much higher than the SPF. Figure (7) shows instead that divergence between the BVAR and the DSGE is due to consumption and output, with the

DSGE forecasting a milder contraction in output in 2021 and a more modest growth in 2022.

While one might expect a large divergence when comparing time series models with experts forecasts, i.e. forecasts based on historical correlations vs more judgment-adjusted, forward-looking forecasts, the large divergence between the two sets of experts' based forecasts in figure (2) and the one between two time series models in (6) are more unexpected. In particular, note that, while we cannot observe the information set available to the SPF and Tealbook, the DSGE and the BVAR forecasts are estimated using a handful of series, mostly the same variables or covering similar concepts, over the same sample, therefore, the discrepancies in the forecasts are mainly due to the cross-equation restrictions imposed by the DSGE model.

The norm rises above the threshold for all recessions, except the first one. This is not surprising: given that the first observations of our evaluation sample coincide with the onset of the 1981-82 recession, the norm fails to detect it, except for the BVAR vs SPF comparison. In those first quarters the norm is high, but falling, as the recession unfolds. This implies that the threshold, which is an average of the past values, lies above the norm. Set aside this episode, the figure highlights that our proposed indicator is quite powerful, as the norm rises above its historical mean during all other recessions. We investigate alternative choices of the threshold in the robustness section.

Wald Test. The Wald test accounts for the correlation across horizons and variables. How does this measure compare to the norm, in its ability to provide early warning signals of turmoil ahead? To answer this question, we plot the time series of the Wald test statistics and p-value for the comparison between the SPF and BVAR forecasts and for the DSGE and BVAR forecasts in Panel (a) and Panel (b) respectively of Figure (9), together with the critical value and the significance level. The variables and forecast horizons included in the Wald test are the same as those included in the norm. The test is computed with information up until the forecast origin. In particular, the covariance matrix V at time t is estimated with vintages of forecasts up to $t-1$. The first t-stat in the sample is computed for the vintage of forecasts made in 1995Q1, and the covariance matrix is estimated over the sample 1981Q3-1994Q4. If the test rejects the null, the forecasts are considered statistically significantly different from each other. Consequently, the Wald test provides a clear cutoff value that can be used to determine whether further examination of the forecasts is necessary. The desired level of "strictness" determines the critical value. Similar to the norm, when examining the time series of both the test-statistics and the p-value of the Wald-test, we observe that the metric spikes during recessions. The pandemic recession is by far the highest spike in both measures, and leads to a rejection of the null in the case of the comparison between the SPF

and the BVAR. Differently from the norm, this measure shows fewer peaks, and no false positive: it only surges during recessions. Furthermore, we note an interesting asymmetry: while the t-stat and p-value increase rapidly in periods of turmoil, they decrease slowly once the recession is over.

Coverage. When full predictive densities or forecast intervals are available, the coverage checks whether one forecast lies within another's fan chart. We obtain 68% confidence intervals from the BVAR model and compute the coverage over the out-of-sample period 1981Q3-2023Q1 for the SPF and BVAR and the DSGE and the BVAR respectively. The time series of the coverage in figures (10) and (11) represents the share of variables and forecast horizons for which the SPF forecasts or the DSGE forecasts are within the 68% credible set of the BVAR forecasts. Note that we cannot compute the coverage for the comparison between SPF and Tealbook forecasts, as interval forecasts for these sets of forecasts are not available.[11] As before, for the SPF vs BVAR comparison, the variables considered are real output, GDP deflator, real consumption, real nonresidential fixed investment, real residential fixed investment and t-Bill over the forecast horizons $h = 1, ..4$. The DSGE vs BVAR comparison excludes the residential and non residential investment. We suggest to compare the coverage to the chosen significance level, illustrated by the black line. The policymaker can thus use this as a sign to further investigate the deviating forecasts.

Like in the case of the norm, we observe substantial deviations of the measure over time. Both figures show that the coverage falls below the significance level during every recession in our sample. We only have two false positive episodes for the SPF vs BVAR comparison, in 1985 and 1986, and a few more for the DSGE vs BVAR. The discrepancies could also be due to the relatively small estimation sample for the BVAR model, which could have led to poor estimation of the predictive density and therefore of the highest posterior density intervals. In fact, in the latest part of the sample, after the Great Financial Crisis, we do not observe false positives. How to determine the reasons behind the low coverage? One approach is to plot the credible intervals, together with the point forecasts of the alternative models, to uncover for which variables and horizons the forecasts fall outside the forecast intervals. Running this exercise, we find that the low coverage in the early 1990 and the 2001 recessions is mainly driven by output and consumption, while during the Great Financial Crisis, by real consumption and real residential investment and the t-Bill rate. During the pandemic, instead, all variables contribute to the low coverage at all forecasting horizons.

[11] One way to circumvent this issue would be to compute a proxy of the interval forecasts using the desired lower and upper percentile of the distribution of point forecasts for the SPF.

4.3 Robustness Checks

We check the robustness of our results through several exercises. First, we evaluate the robustness of the norm to outlier observations, defined as abnormally high discrepancies in a given forecast vintage. To do so, we recompute the norm windsorizing extreme gaps. Specifically, at each time t in our sample, we set the values of the L-th largest standardized squared deviations to the value of the largest L-th-1 deviation. We have a relatively small number of deviations for every time t, given by NK , corresponding to 20 for the SPF vs Tealbook and BVAR vs DSGE comparison, and 24 for the BVAR vs SPF one. Therefore, we set $L = 2$ for the first two and $L = 3$ for the latter. Figure (12) shows that even when windsoring the outliers, we get very similar spikes to our baseline computation of the norm for all comparisons. This suggests that surges in the norm are driven by several large gaps occurring at the same forecast origin.

Second, we explore the sensitivity of the norm to the choice of the threshold. In our baseline analysis, we use a rolling mean with window of 40 observations. First, we consider a different window size, 20 observations, to account for short history of forecasts and/or frequent structural breaks. Using a shorter window size increases the likelihood of anticipating a recession, as shown in figure (13). We then run an additional exercise, where we compute the True Positive Rate (TPR) and the False Positive Rate (FPR) for different window sizes from $w = 2, .., 140$. We define a true (false) positive a situation in which the norm is above the rolling mean of size w and a recession occurs (does not occur) within the next four quarters. This way, we assess the ability of the norm to provide early warnings of severe economic turmoil ahead. As shown in figure (14), we find that both the TPR and FPR are generally declining in the window size. This indicates that, usually, the larger the window size, the smaller is the threshold. Therefore, one should prefer to use a smaller window size if concerned about failing to detect a recessionary event. As an alternative approach to the historical mean, we define the threshold at time t as the 75th percentile of the norm distribution observed up to time $t - 1$. Figure (15) highlights that this approach would lead to miss not only the 1981-82 for all comparisons, but also the 1990 recession. Additionally, we define as alarming events, periods when the norm is increasing for two consecutive quarters. As in the case of the definition based on percentiles, we fail to detect the 1990 recession, see figure (16). Overall, we recommend using a rolling mean as the threshold due to its straightforward implementation and ability to flag turbulence ahead.

The estimation of the variance-covariance matrix V in the implementation of the Wald-test can be problematic in environments when the sample size is limited compared to the number of variables and horizons included in the comparison. The use of shrinkage or factor-based estimators could mitigate the problem. As an example, we use the linear shrinkage

estimator proposed by Ledoit and Wolf (2004), which is a convex linear combination of the sample covariance matrix with the identity matrix. We apply this estimator in the case of the SPF and BVAR comparison, where the number of variables is the largest, so that $NK = 20$. Unsurprisingly, figure (17) shows that in this case the t-statistics becomes more similar to the norm.

Finally, we repeat the coverage exercise using different credible bands to show how the choice of fan-chart width affects the informativeness of this metric. We find that using a 90% significance level still provides signals of large deviations for both comparisons during all recessionary episodes, while using the more stringent 95% level the coverage measure detects significant deviations between the SPF and BVAR models, but does not uncover gaps during the recession associated with the COVID-19 pandemic, as shown in figure (18). Therefore, as expected, the coverage increases with the width of the confidence intervals, but, less trivially, our measure still detects significant events of downturn, even with very wide bands.

In our baseline analysis, we consider forecasts horizons from nowcast till the four-quarter-ahead due to the availability of SPF data. However, central banks might care about longer forecasting horizons. Because of the lags in the transmission channel of monetary policy, the horizons six to eight quarters ahead might be more relevant for monetary policy authorities. Therefore, we conduct the BVAR and DSGE comparisons at these horizons to check whether the metrics retain their real-time alert properties. Figure (19) highlights that looking at longer horizons actually enhances the ability of the norm to anticipate recessions, particularly the U.S. recession of 1990-1991 and the Great Recession, that were almost undetected by the norm applied to shorter horizons. The Wald test never rejects the null hypothesis, as in the baseline analysis, and the Wald-statistic larger spikes coincide with the 2001 and the Great Recession, rather than with the recession associated with the COVID-19 pandemic. This probably reflects large difference in the covariance matrix of shorter vs longer horizons differentials. Finally, similarly to the norm, the coverage provides stronger signal of troubles ahead when applied to long, rather than short horizon forecasts.

4.4 Cross-Check Algorithm

With the insights offered from our empirical results and robustness checks, we suggest the following simple approach when conducting a cross-check evaluation, outlined in these steps:

1. **Select the preferred models or set of forecasts** to be included in the cross-check.[12]

[12]Though a joint cross-check with several models would be interesting it is outside the scope of the paper.

Choosing an accurate alternative model can shield against large, systematic forecast errors. This can be done by comparing the relative forecasting performance of a model against possible alternatives, if the objective of the cross-check is to avoid large forecast errors and ultimately, large policy mistakes.

2. **Pick the preferred measures of discrepancy.** This can be determined based on the information available and the preferences of the policy maker, keeping into account the following: (i) The normalized Euclidean norm is the appropriate measure for a practitioner seeking a simple indicator of the most extreme outliers, as it picks out the largest standardized deviations. (ii) For early detection of widespread model breakdowns, the Wald test delivers p-values that consider cross-correlations. Implementation, however, is infeasible with small sample sizes—especially when smaller than the product of the number of observations and the number of forecast horizons. (iii) When full predictive densities are at hand, use a coverage measure to verify whether one forecast falls inside another's confidence intervals.

3. **Compute the measures.** In order to do that, several important parameters and implementation details should be selected: the significance level for the Wald and coverage tests, the evaluation sample for the estimation of the variance-covariance matrix in the Wald test, the threshold value for the norm, the variables and horizons to include in the cross-check, as well as the weighting scheme, the treatment of outliers. Our robustness checks provide some guidance: a rolling mean of the latest 10 years of data is a simple and informative threshold for the norm, flexible enough to account for structural breaks and powerful enough to detect future downturns. A significance level of up to 90% for the coverage seem to be high enough to detect periods of turmoil. Windsorizing the largest discrepancies preserves the informativeness of the norm. When longer forecast horizons are available, they should be included in the comparison, as they seem to better anticipate recessions than short term horizons.

4. **Compare measures with threshold values**, i.e. determine whether the value taken by the measure supports the hypothesis that the forecasts are diverging too much. This can be done based on the significance levels selected in Step 3, for the Wald test and the coverage, and based on the pre-set threshold for the norm.

5. **Analyze the individual forecasts** for each variable and horizon, in case the criterion set above requires further investigation. Plotting the standardize square differentials as in figures (4), (7) or (8) or the confidence intervals together with the alternative model point forecasts can identify the variables contributing to the rejection of the

null of no divergence. Then, the reasons for the divergence, e.g. initial conditions and cross-equation restrictions, can be determined.

5 Conclusion

We suggest several measures to compare the forecasts from two different models or sets of forecasts. The measures summarize distance *jointly* across variables and horizons. We illustrate the usefulness of our measures when comparing the forecasts made by experts, i.e. SPF and Tealbook, or obtained from time series models, a medium scale BVAR and a DSGE model with U.S. data. The models forecasts depart during recessions, making our measure spike. These measures complement, rather than substitute, forecast accuracy evaluation. The cross-checks proposed in this paper are not merely academic; they are highly policy-relevant for central banks regardless of whether they rely on a single "core" model or have a systematic use of model suites with reliance on survey based forecasts. As a result, real-time cross-check metrics are essential safeguards for robust policy deliberation.

References

Brandao-Marques, L., G. Gelos, D. Hofman, J. Otten, G. K. Pasricha, and Straussoe (2024). Do household expectations help predict inflation? Technical Report DP18774, International Monetary Fund Discussion Paper Series.

Capistrán, C. (2006). On comparing multi-horizon forecasts. *Economics Letters 93*(2), 176–181.

Christoffersen, P. F. (1998). Evaluating inteval forecasts. *International Economic Review 39*(4), 841–862.

Clark, T. E. and M. W. McCracken (2012). Reality Checks and Comparisons of Nested Predictive Models. *Journal of Business and Economic Statistics 30*(1), 53–66.

Crushore, D. and T. Stark (2019). Fifty years of the survey of professional forecasters. *Economic Insights 4*(4), 1–11.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*(3), 253–63.

Faust, J. and J. H. Wright (2013). *Forecasting inflation* . Handbook of Economic Forecasting: Elsevier.

Federal Reserve Bank of Philadelphia (2024a). Survey of professional forecasters: Median forecast in levels. https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters. Accessed: 2024-11-30.

Federal Reserve Bank of Philadelphia (2024b). Tealbook data set. https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/philadelphia-data-set. Accessed: 2024-11-30.

Federal Reserve Bank of St. Louis (2024). Fred economic data. https://fred.stlouisfed.org. Accessed: 2024-11-30.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Giannone, D., M. Lenza, and G. E. Primiceri (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics 97*(2), 436–451.

Granziera, E., K. Hubrich, and H. R. Moon (2014). A Predictability Test for a Small Number of Nested Models. *Journal of Econometrics 182*, 174–185.

Granziera, E., V. Larsen, G. Meggiorini, and L. Melosi (2025). Speaking of Inflation: The Influence of Fed Speeches on Expectations. Technical Report 20038, CEPR Discussion Paper.

Granziera, E. and T. Sekhposyan (2019). Predicting relative forecasting performance: An empirical investigation. *International Journal of Forecasting 35*, 1636–1657.

Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business Economic Statistics 23*(4), 365–380.

Hansen, P. R. (2011). The Model Confidence Set. *Econometrica 79*(2), 453–497.

Hoesch, L., B. Rossi, and T. Sekhposyan (2023). Has the information channel of monetary policy disappeared? revisiting the empirical evidence. *American Economic Journal: Macroeconomics 15*(3), 355–387.

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

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Lenza, M. and G. E. Primiceri (2022). How to Estimate a VAR after March 2020. *Journal of Applied Econometrics 37*(4), 688–699.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *The Review of Economics and Statistics 69*(4), 667–674.

Norges Bank (2022). Norges bank's monetary policy handbook. Accessed: 2024-11-29.

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

Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business Economic Statistics 39*(1), 40–53.

Sims, C. (2002). The Role of Models and Probabilities in the Monetary Policy Process. Technical Report 2:2002, Brooking Papers on Economic Activity.

Smets, F. and R. Wouters (2007). Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review 97*(3), 586–606.

Tsiaplias, S. (2020). Time-varying consumer disagreement and future inflation. *Journal of Economic Dynamics and Control 116*, 1–18.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica 68*(5), 1097–1126.

Table 1: Summary of Measures

	Norm	**Wald-Stat Model 1**	**Wald-pvalue Model 2**	**Coverage**
Support	$[0, \infty)$	$[0, \infty)$	$[0, 1)$	$[0, 1)$
Sample	1	$P >> KH$		1
Parameter		α		α
Information	point	point		interval
Correlation	no	yes		no
Caveats	outliers threshold	Gaussian assumption estimation of V		correct density specification choice of interval width
FA Analogue	multivariate RMSFE	Diebold and Mariano (1995) & Capistrán (2006) tests		Christoffersen (1998) interval test

Note. "Support": range of values for the measure. "Sample": minimum number of forecast vintages needed to compute the measure. "Parameter" α: confidence level of the test for Wald-test, confidence level of the forecast interval for the coverage. "Information": type of forecast, only point for the norm and Wald test, interval of at least one model for the coverage. "Correlation": indicates whether the measure takes into account the correlation among variables and forecasting horizons. "Caveats": issues associated with the measure: norm's sensitivity to outliers and threshold choice; Wald's estimation of the variance covariance matrix V and Gaussian assumption; coverage's reliance on correct density specification and choice of interval width. "Forecast Accuracy Analogue": closest forecast evaluation counterpart of the cross-check measure.

Table 2: Relative RMSFE: Alternative Forecasts vs SPF

	h=1	**h=2**	**h=3**	**h=4**
			BVAR	
RGDP	1.38**	1.35***	1.27**	1.08
PGDP	1.25**	1.29**	1.35**	1.39**
RRESINV	1.19*	1.13*	1.13*	1.17*
RCON	1.27*	1.33	1.29	1.17*
TBILL	1.94***	1.67***	1.51***	1.42***
			DSGE	
RGDP	1.22**	1.30***	1.33***	1.28***
PGDP	1.07	1.07	1.02	1.00
RCON	1.26*	1.36***	1.36***	1.31*
TBILL	1.88***	1.49***	1.27***	1.10*
			TealBook	
RGDP	0.80**	0.98	0.99	0.99
PGDP	0.98	0.98	0.94	0.88*
RRESINV	0.86*	0.85*	0.83*	0.89
UNEMP	0.93	0.91*	0.91*	0.93

Note: Relative RMSFE with respect to the SPF forecasts, over the sample 1981Q3-2023Q1. Forecasts are evaluated against the final release (2025Q1 vintage). '*', '** and '***' indicate the significance levels for the equal predictive ability test by Diebold and Mariano (1995) at the 10, 5 and 1 percent respectively.). The last observation for the Tealbook forecasts is 2019Q4.

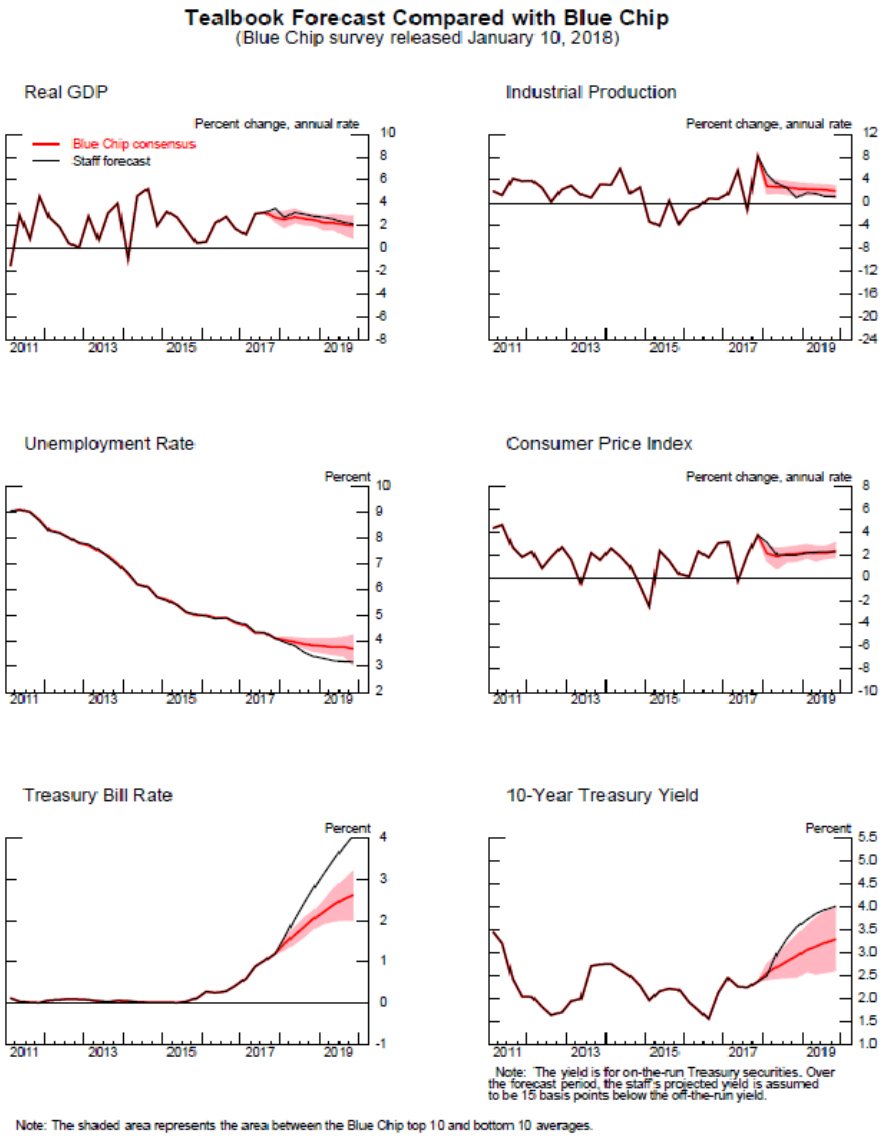


Figure 1: Tealbook vs Blue Chip survey forecasts, January 2018.

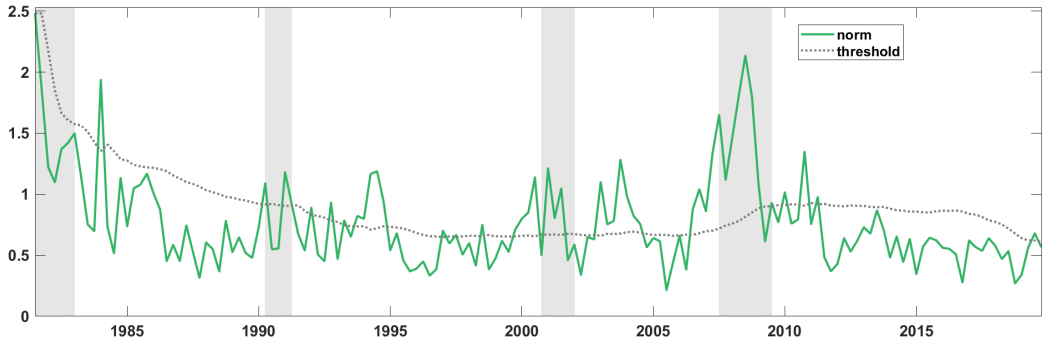
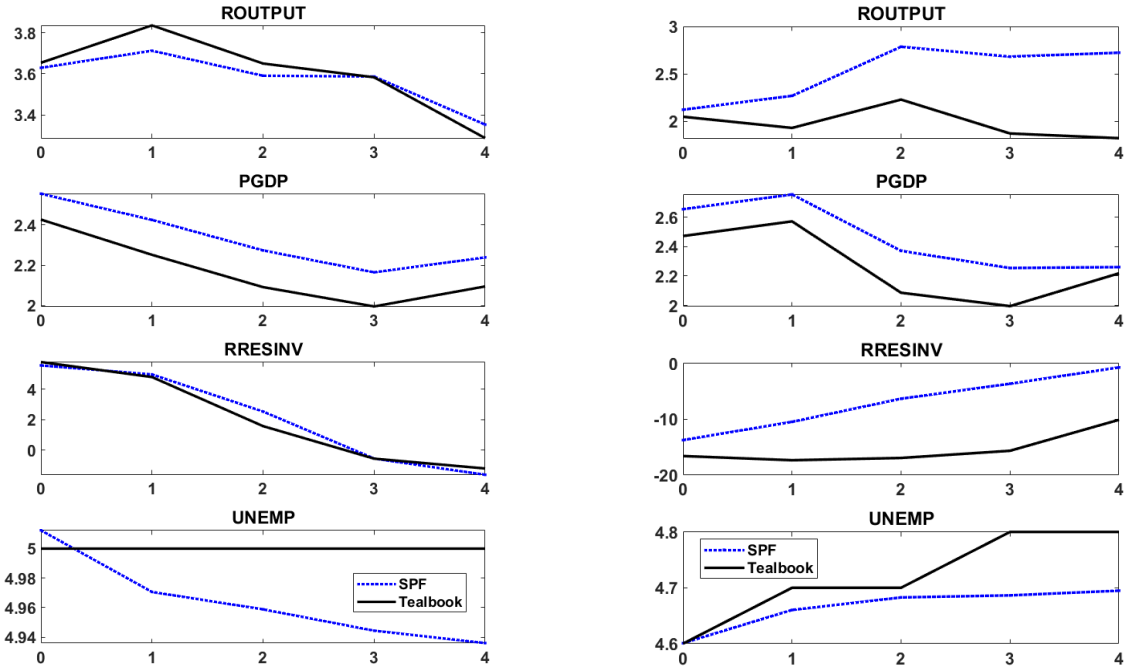


Figure 2: Norm (green, solid line) for SPF and Tealbook forecasts and its rolling mean (black, dashed line) over a ten year window. Grey shaded areas denote the NBER recessions. The norm is computed over real output, PGDP inflation, real residential investment and the unemployment rate over the forecast horizons $h = 0, ..4$. Sample 1981Q3-2019Q4.



(a) Forecasts made in 2005Q3

(b) Forecasts made in 2007Q3

Figure 3: SPF (blue dotted) and Tealbook (black solid) forecasts for real output, PGDP inflation, fixed residential investment and unemployment rate over the forecast horizons $h = 0, ..4$.

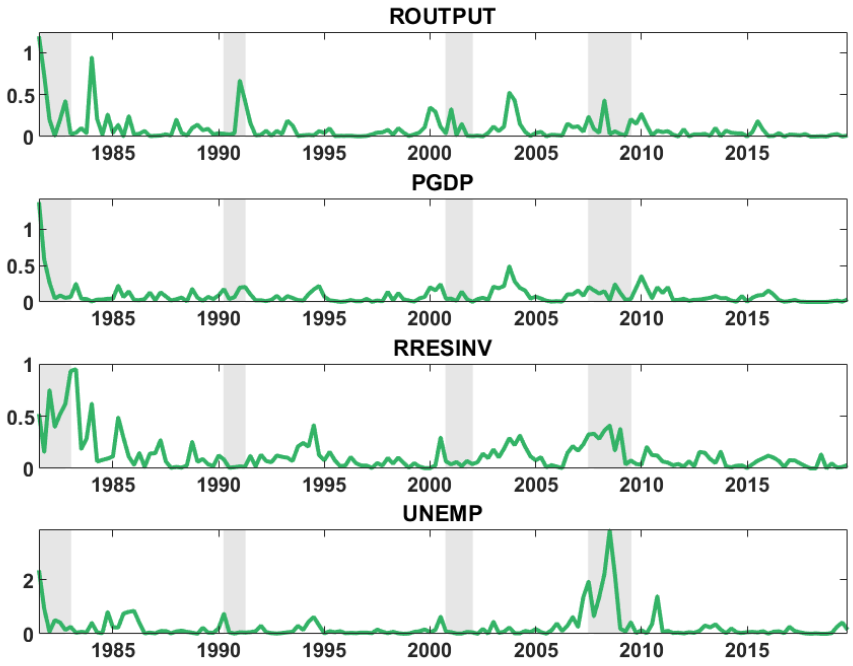


Figure 4: Time series of squared deviations between the SPF and Tealbook forecasts described in section (3) for real output, GDP deflator, real residential investment and the unemployment rate computed over the forecast horizons $h = 1, ..4$, sample 1981Q3-2019Q4. Grey shaded areas denote the NBER recessions.

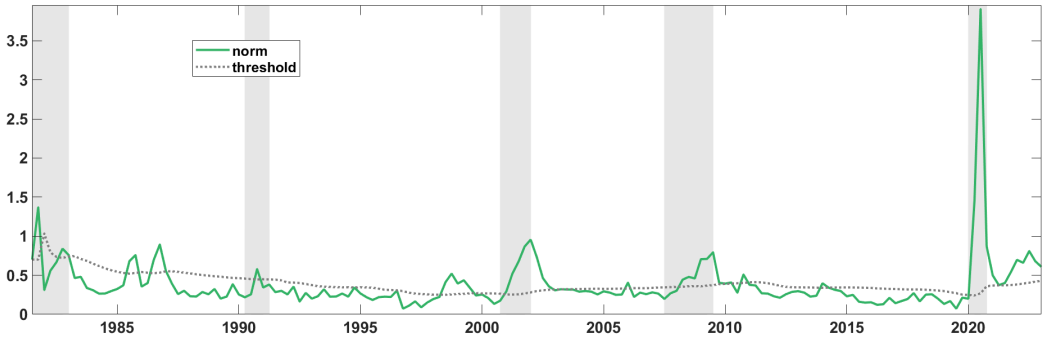


Figure 5: Norm (green, solid line) and threshold (black, dashed line) for the SPF and the medium scale BVAR model described in section (3) for real output, GDP deflator, real consumption, real residential fixed investment, real non residential fixed investment and t-Bill over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions. Time series for the sample 1981Q3-2023Q1.

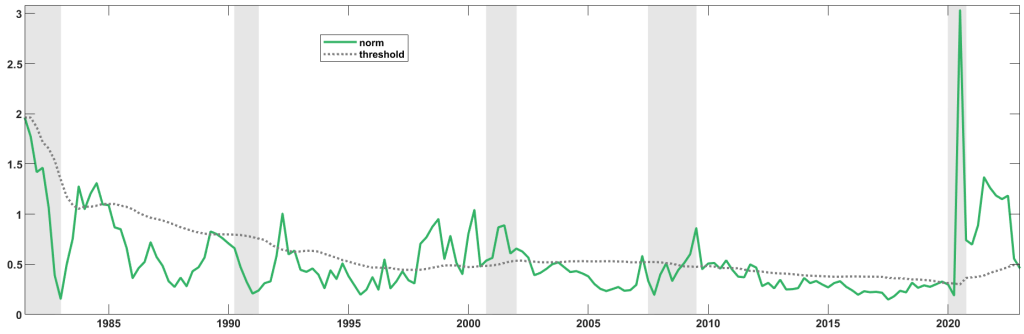


Figure 6: Norm (green, solid line) and threshold (black, dashed line) for the medium scale BVAR and DSGE models described in section (3) computed over real output, GDP deflator, real consumption and the t-Bill rate over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions. Time series for the sample 1981Q3-2023Q1.

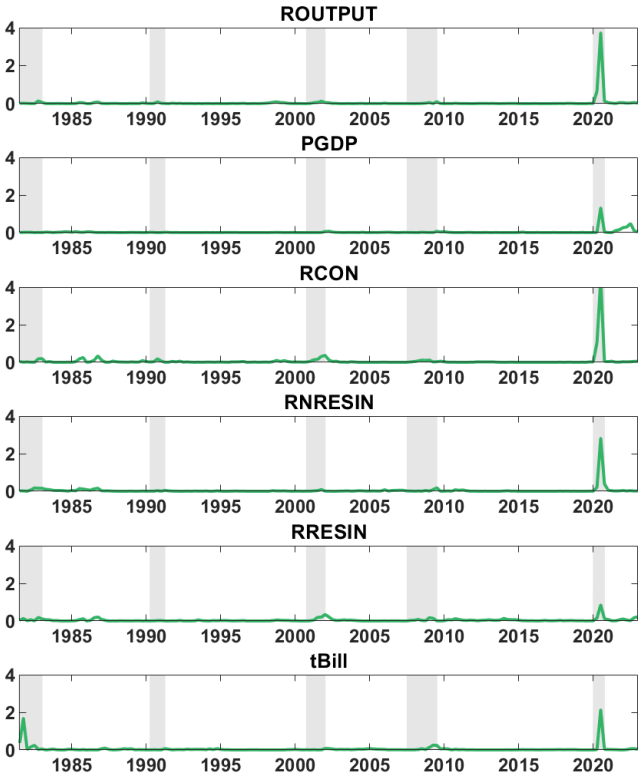


Figure 7: Time series of squared deviations between the SPF and BVAR models described in section (3) for real output, GDP deflator, real consumption, real residential investment, real non-residential investment and the t-Bill rate computed over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions. Time series for the sample 1981Q3-2023Q1.

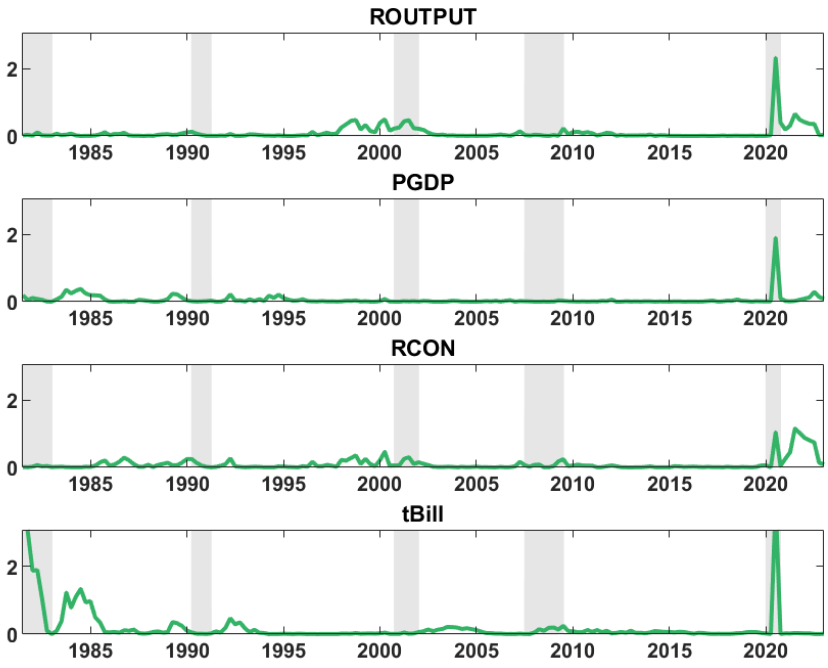
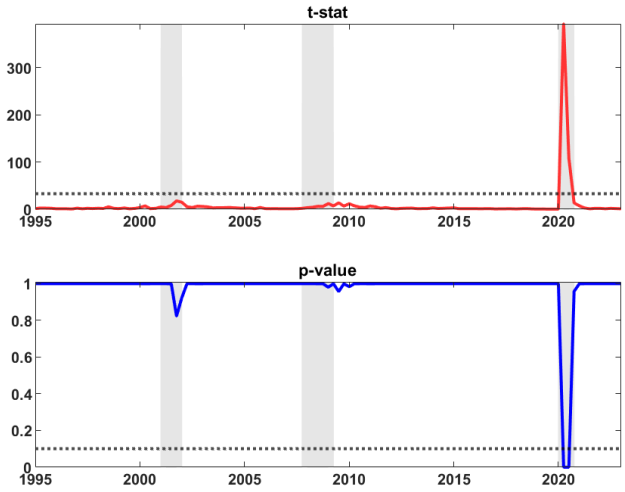
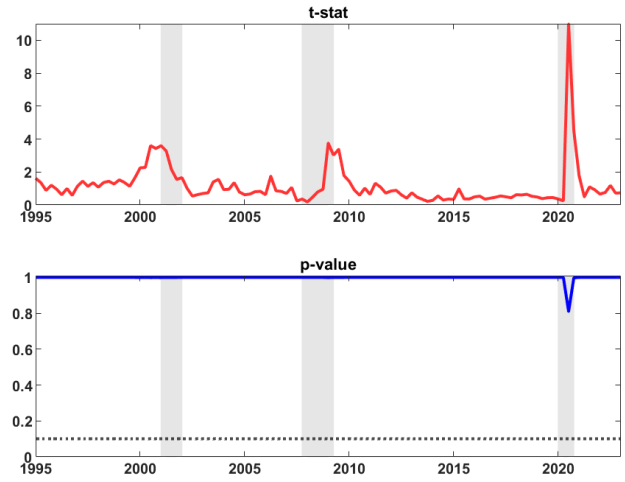


Figure 8: Time series of squared deviations between the BVAR and DSGE models described in section (3) for real output, GDP deflator, real consumption and the t-Bill rate computed over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions. Time series for the sample 1981Q3-2023Q1.



(a) SPF vs BVAR comparison



(b) DSGE vs BVAR comparison

Figure 9: Wald test at the 90% confidence level, time series for the sample 1995Q1-2023Q1. Panel (a): SPF and BVAR forecasts of real output, GDP deflator, real consumption, real nonresidential fixed investment, real residential fixed investment and t-Bill over the forecast horizons $h = 1, ..4$. Panel (b): DSGE and BVAR forecasts of real output, GDP deflator, real consumption and the t-Bill over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions.

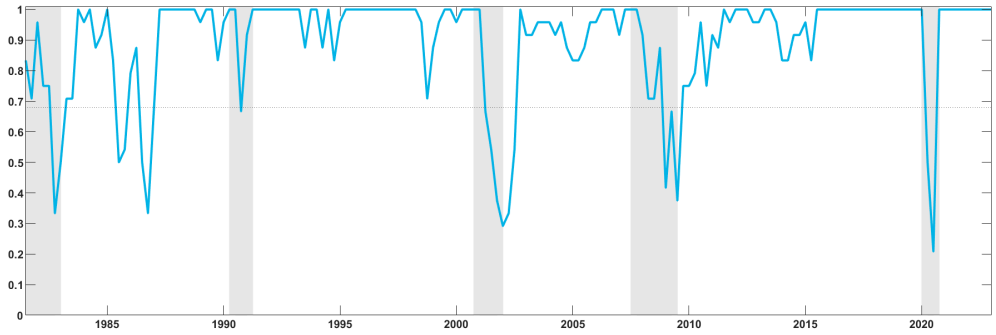


Figure 10: Coverage (68%) (blue, solid line) and significance level of confidence intervals (black, dashed line) for SPF and medium size BVAR over real output, GDP deflator, real consumption, real nonresidential fixed investment, real residential fixed investment and t-Bill over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions. Time series for the sample 1981Q3-2023Q1.

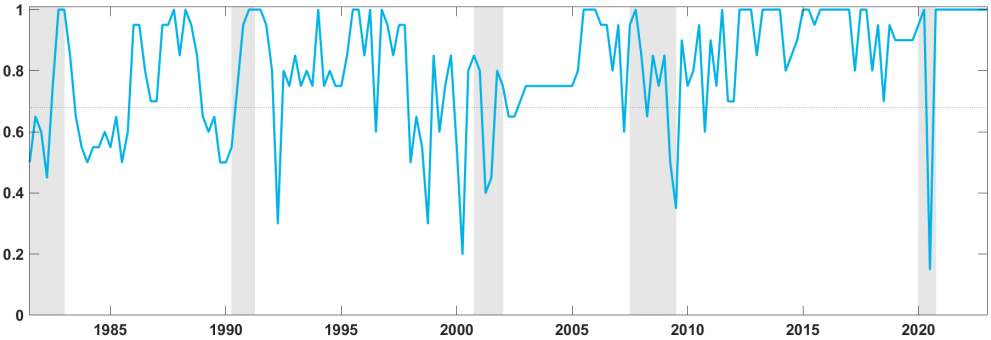
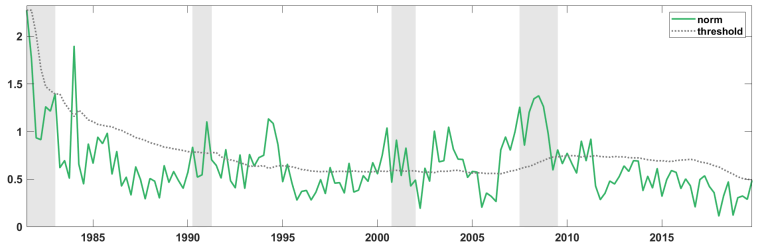
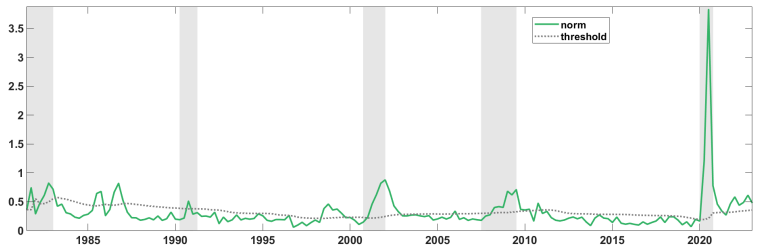


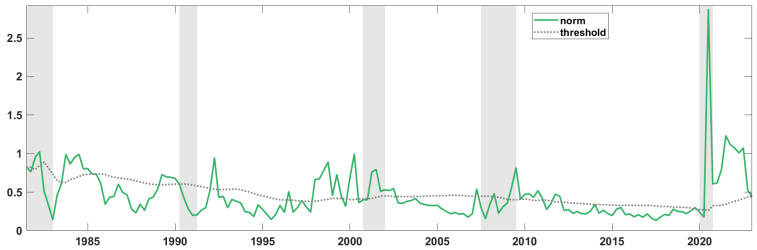
Figure 11: Coverage (68%) (blue, solid line) and significance level of confidence intervals (black, dashed line) for DSGE and medium size BVAR over real output, GDP deflator, real consumption, and t-Bill over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions. Time series for the sample 1981Q3-2023Q1.



(a) SPF and Tealbook comparison

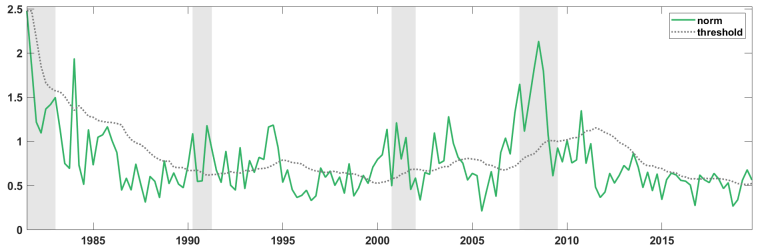


(b) SPF and BVAR comparison

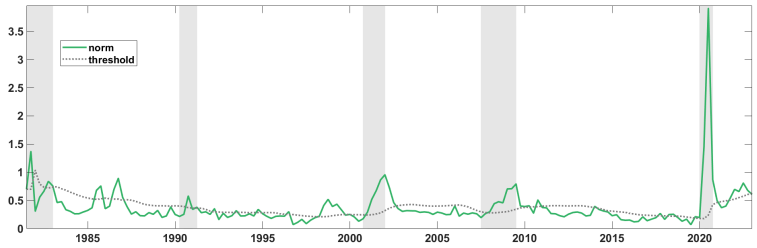


(c) DSGE and BVAR comparison

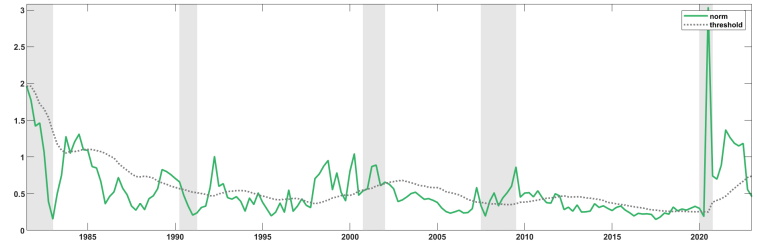
Figure 12: Norm (green, solid line) windsorized and threshold (black, dashed line): two (three) largest values for the SPF vs Greenbook and BVAR vs DSGE (SPF vs BVAR) comparison are set to the third (fourth) largest value for BVAR vs SPF (top panels) and BVAR vs DSGE (bottom) comparisons. Grey shaded areas denote the NBER recessions.



(a) SPF and Tealbook comparison

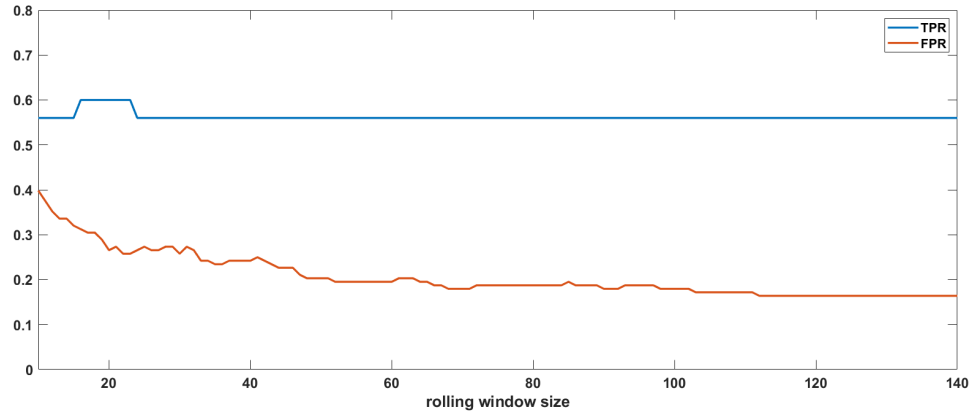


(b) SPF and BVAR comparison

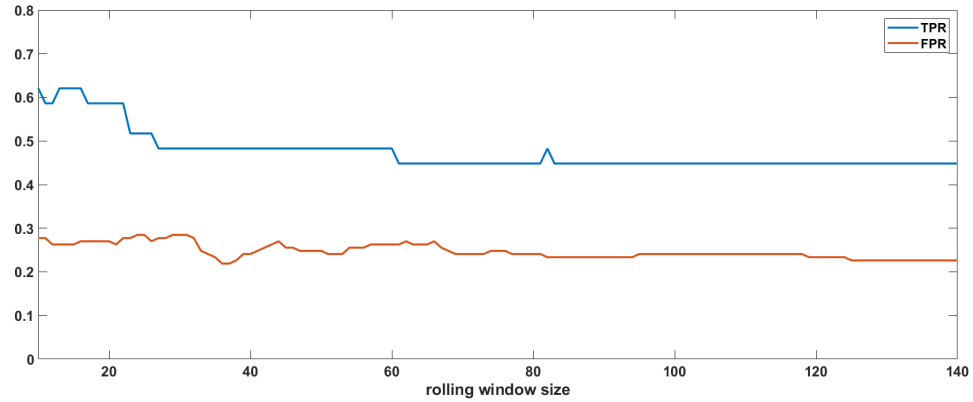


(c) DSGE and BVAR comparison

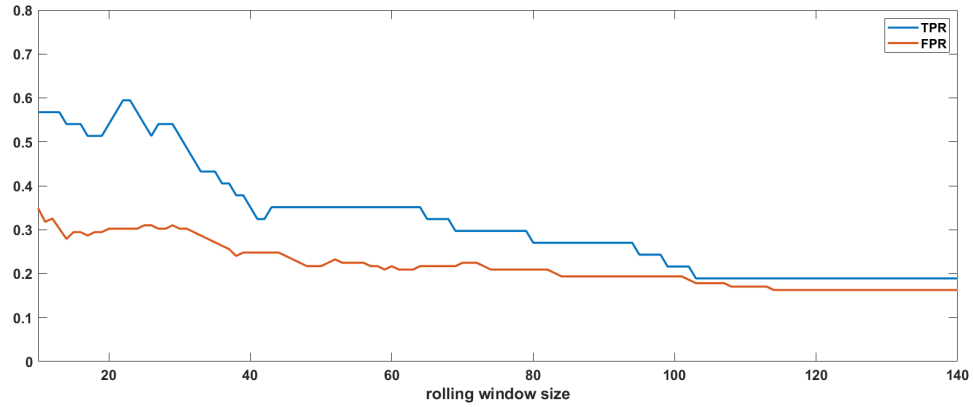
Figure 13: Norm (green, solid line) and rolling mean (black, dashed line), window size= 20: two (three) largest values for the SPF vs Tealbook and BVAR vs DSGE (SPF vs BVAR) comparison are set to the third (fourth) largest value for BVAR vs SPF (top panels) and BVAR vs DSGE (bottom) comparisons. Grey shaded areas denote the NBER recessions.



(a) Norm

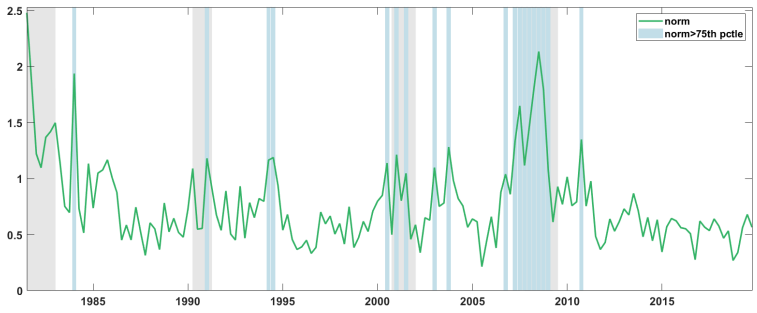


(b) Wald test

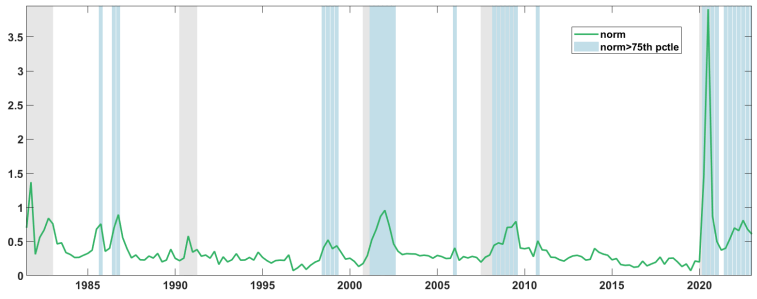


(c) Wald test

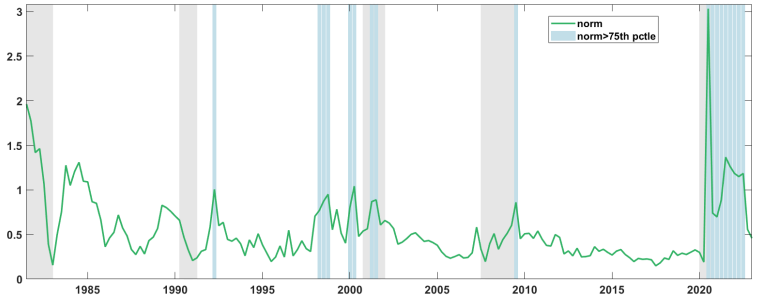
Figure 14: True Positive Rate (blue) and False Positive Rate (orange) for the SPF vs Tealbook, BVAR vs SPF and BVAR vs DSGE comparisons.



(a) SPF and Tealbook comparison

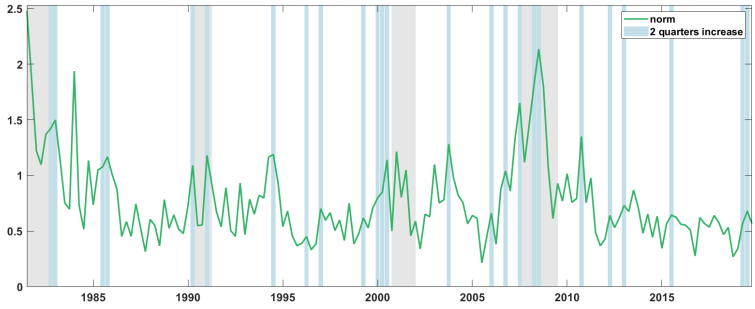


(b) SPF and BVAR comparison

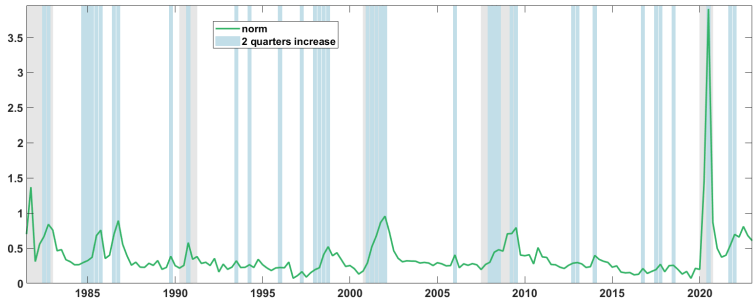


(c) DSGE and BVAR comparison

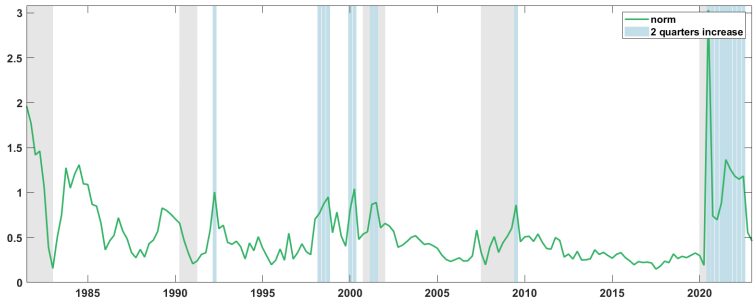
Figure 15: Norm (gree, solid line) and norm above the 75th percentile (blue, shaded areas): SPF vs Tealbook (top panel), SPF vs BVAR (middle panel) and BVAR vs DSGE (bottom panel) comparisons. Grey shaded areas denote the NBER recessions.



(a) SPF and Tealbook comparison



(b) SPF and BVAR comparison



(c) DSGE and BVAR comparison

Figure 16: Norm (green, solid line) and two consecutive quarters of increase in the norm (blue, shaded areas): SPF vs Tealbook (top panel), SPF vs BVAR (middle panel) and BVAR vs DSGE (bottom panel) comparisons. Grey shaded areas denote the NBER recessions.

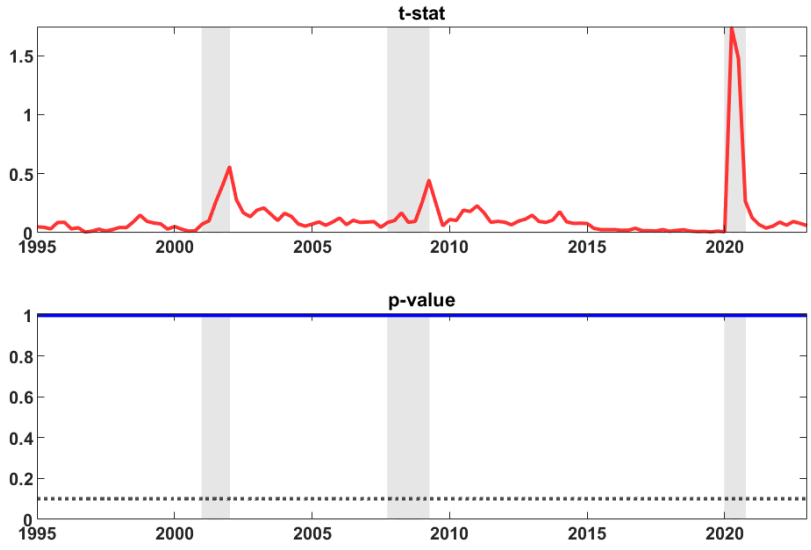
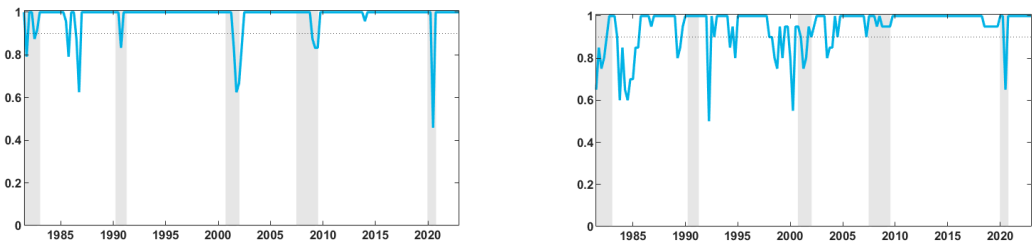
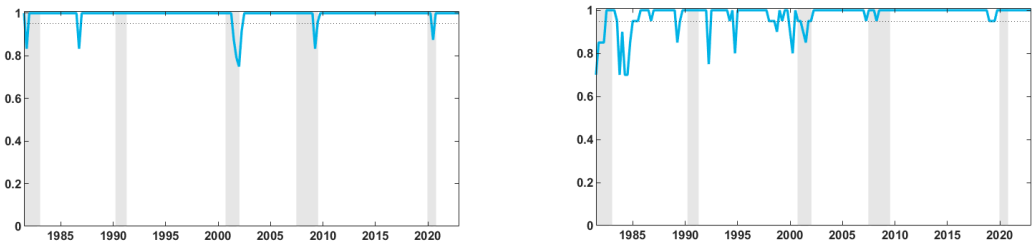


Figure 17: Wald test with shrinkage variance-covariance matrix estimator for the SPF and the BVAR models described in section (3) over real output, GDP deflator, real consumption, real nonresidential fixed investment, real residential fixed investment and t-Bill over the forecast horizons $h = 1, ..4$. Grey shaded areas denote the NBER recessions. Time series for the sample 1995Q1-2023Q1.

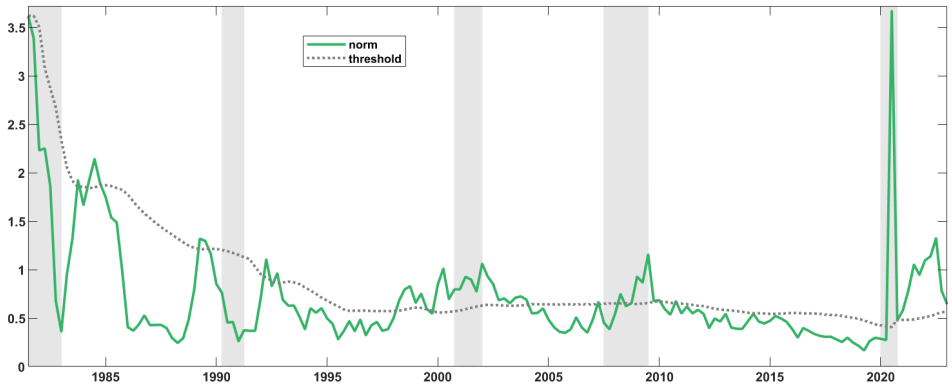


(a) 90% coverage

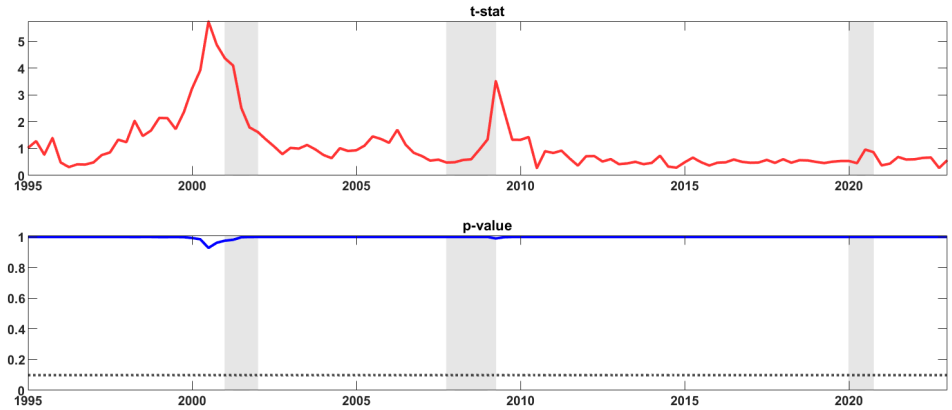


(b) 95% coverage

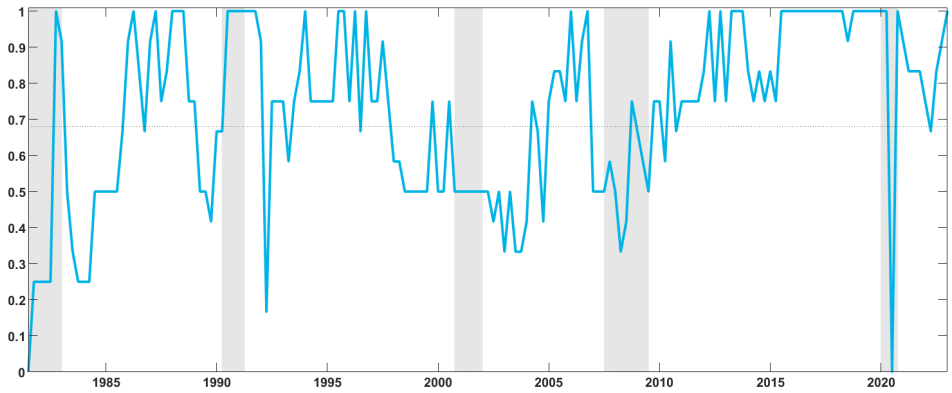
Figure 18: Coverage (blue, solid line): 90% (top panels), 95% (bottom), for BVAR vs SPF (left panels) and BVAR vs DSGE (right) comparisons. Grey shaded areas denote the NBER recessions.



(a) Norm



(b) Wald test



(c) Coverage

Figure 19: BVAR vs DSGE comparisons at longer forecasting horizons, $h = 6, 7, 8$. Grey shaded areas denote the NBER recessions.