A comprehensive evaluation of macroeconomic forecasting methods[*]

Andrea Carriero
Queen Mary University of London

Ana Beatriz Galvao[†]
University of Warwick

George Kapetanios
King's College London

January, 2016

Abstract

This paper contributes to the academic literature and the practice of macroeconomic forecasting. Our evaluation compares the performance of four classes of state-of-art forecasting models: Factor-Augmented Distributed Lag (FADL) Models, Mixed Data Sampling (MIDAS) Models, Bayesian Vector Autoregressive (BVAR) Models and a medium-sized Dynamic Stochastic General Equilibrium Model (DSGE). We look at these models to predict output growth and inflation with datasets from the US, UK, Euro Area, Germany, France, Italy and Japan. We evaluate the accuracy of point and density forecasts, and compare models with a large set of predictors with models that employ a medium-sized dataset. Our empirical results shed light on how the predictive ability of economic indicators for output growth and inflation changes with horizon, on the impact of dataset size on the calibration of density forecasts, and how the choice of the multivariate forecasting model depends on the forecasting horizon.

Keywords: factor models, BVAR models, MIDAS models, DSGE models, density forecasts.

JEL codes: C53

[*] All three authors acknowledge support for this work from the Economics and Social Research Council [ES/K010611/1]. We also are grateful to our reseach assistant Katerina Petrova, who is also funded by the ESRC, for excellent assistance.

[†] Corresponding author: Ana Galvao, Economic Modelling and Forecasting Group - Warwick Business School, University of Warwick, Coventry, CV4 7AL, UK. ana.galvao@wbs.ac.uk.

1 Introduction

Forecasting is one of the major aims of economic and econometric analysis along with modelling the foundations of economic phenomena. Arguably, forecasting is one of few activities of academic and professional economists which is an end to itself, in the sense that attempting to know what the future holds is a universal human endeavour. As a result, considerable efforts have been made in academic work to lay the foundations and build tools for efficient forecasting. That work has had to face inevitable difficulties and criticisms given the challenge of the task at hand. In particular macroeconomic forecasting has long been viewed as a suspect activity given the potential for forecasting models to exhibit spuriously good performance that subsequently fades away when the model is used repeatedly in action.

The macroeconomic forecasting literature can be divided into two large parts. The first aims to produce models that attempt to explain the economy first and then provide forecasts only as a byproduct of their main aim. This is, in principle, optimal in the sense that a model which can explain successfully the economy should be able to forecast well. Nevertheless the complexity of the economy and of the models that are needed for its full explanation implies that such forecasts might not be accurate in sample, let alone out of sample.[1] The second considers models that do not attempt a full structural modelling but simply a reduced-form statistical description. These models frequently have superior forecasting performance, but their reduced-form nature coupled with their volatile performance imply that they are viewed with suspicion by economists and policymakers.

This has not stopped the proliferation of reduced-form models and a rapid rise in their sophistication. Recent trends in this literature include modelling structural changes and the efficient use of increasingly larger datasets. The former has been driven by the widespread recognition that structural change is a leading cause of forecast failure. A number of approaches of varying sophistication are being used to accommodate structural change. These range from time-varying coefficient models to methods that allow for time varying estimation of standard econometric forecasting models. In this context, as is common with forecasting in general, increasing sophistication has not been found to correlate closely with superior forecasting performance.[2] The second trend of considering large

[1] For example, Faust and Wright (2013) and Chauvet and Potter (2013) conclude their reviews on the forecasting performance of structural and reduced-form models for predicting inflation and output growth arguing that structural models do not have better forecast accuracy than univariate time series models.

[2] For example, Faust and Wright (2013) provide evidence that time-varying vector autoregressive models with stochastic volatility do not improve point forecasts of inflation in comparison with a univariate benchmark, although there is stronger evidence that stochastic volatility improves density forecasts of inflation (Clark, 2011). Chauvet and Potter (2013) consider Markov-Switching models to predict output growth, and they find gains only during recessions and only at short horizons. Based on data for a set of countries, Ferrara, Marcellino and Mogliani (2015) show that nonlinear models rarely improve forecasts of their linear counterpart.

datasets has been spurred by their use in many economic analyses, given their availability in central banks and other policy making institutions.[3]

The above developments set the scene for the current paper. Our aim is to provide a state of the art and comprehensive evaluation of recently proposed forecasting model classes, giving special attention to model classes able to deal with large datasets. We assess the same classes of models and forecasting horizons for predicting output growth and inflation so we can evaluate whether it is adequate to use one forecasting model to predict these two popular macroeconomic variables. The scope of the paper is to strike a balance between being comprehensive and producing clear messages. This requires considering a wide range of models but being selective in some dimensions so as to make the evaluation exercise feasible and informative. Further, it requires an evaluation across a number of different countries and different sample periods. Finally, we aspire to compare and contrast reduced-form models and structural models, which have traditionally been considered inferior for forecasting purposes. This later aspect of our analysis is less commonly found in forecasting evaluations. [4]

Our forecasting exercise compares the forecasting performance of four classes of state-of-art forecasting models: Factor-Augmented Distributed Lag (FADL) Models, Mixed Data Sampling (MIDAS) Models, Bayesian Vector Autoregressive (BVAR) Models and a medium-sized Dynamic Stochastic General Equilibrium Model (DSGE). We have knowledge of the relative forecasting performance of DSGE models with respect to Bayesian VARs (as, for example, Smets and Wouters (2007)), of FADL to Factor-Augmented MIDAS Models (Andreou, Ghysels and Kourtellos (2013)), and of Bayesian VARs to Dynamic Factor models (Bańbura, Giannone and Reichlin (2010)). However, in this paper we look at all these models to predict output growth and inflation. Forecasting comparisons in the literature focus normally on data from a single country or a small subset of countries (US, UK and Euro Area).[5] We will use data from seven economies: US, UK, Euro Area, Germany, France, Italy and Japan. Our reduced-form forecasting models —Factor, BVAR and MIDAS— are useful to exploit the predictive information on large datasets so we built a large dataset for each one of these countries. We will also assess the importance of employing large (one-hundred predictors)

[3] Stock and Watson (2002) is an influential paper supporting the use of large datasets for forecasting macroeconomic variables.

[4] Density forecasts of DSGE models are evaluated by Del Negro and Schorftheide (2013), but when DSGE models are compared with a large set of statistical models in Faust and Wright (2013) and Chauvet and Potter (2013) only point forecasts are considered. Note also that the set of forecasting models for predicting inflation in Faust and Wright (2013) differs from the models in Chauvet and Potter (2013). While Faust and Wright (2013) consider up to one-year-ahead horizons, Chauvet and Potter (2013) choose to look at horizons up to two quarters only, but Del Negro and Schorftheide (2013) evaluate horizons up to two years ahead.

[5] Stock and Watson (2003) and Kuzin, Marcellino and Schumacher (2013) are exceptions by considering data from seven countries when designing their forecasting exercises. Ferrara, Marcellino and Mogliani (2015) evaluate models for 19 countries, but they use only a relatively small set of predictors.

datasets in comparison with medium-sized (a dozen predictors) and small datasets in macroeconomic forecasting.

We find that multivariate models with either small or large set of predictors do not perform better than univariate models if the target is output growth in two years. This lack of predictive power of economic indicators for output growth at the two-year horizon is not found when predicting inflation. Dissimilarities in the degree of predictability of output growth and inflation were first reported with US survey data (Lahiri and Sheng, 2010; Patton and Timmermann, 2011). The medium-scale DSGE model (Smets and Wouters, 2007) provides accurate forecasts of US and UK inflation at long horizons, expanding the results of Del Negro and Schorftheide (2013) with US data.

The choice of the best model class for macroeconomic forecasting depends on the forecasting horizon. At the nowcasting horizon (forecasting the current quarter; $h = 1$), mixed frequency models (MIDAS) provide accurate point forecasts of output growth and density forecasts of quarterly inflation, but they fail when predicting annual (change on the same quarter a year ago) inflation. At longer horizons, Bayesian VARs and, in some circumstances, factor models are more accurate.

We find no strong support to the use of large datasets (one-hundred predictors) instead of medium (a dozen predictors), but gains from using a large dataset are more frequent in the most recent period (2008-2011). For the US and UK output growth, we find that increasing the size of the dataset improves the calibration of density forecasts.

We describe the classes of forecasting models in Section 2. Section 3 provides a summary of the datasets we employed, which are fully reported in our online appendix. Section 4 describes the design of our forecasting exercise, including statistical tests employed and how we address issues of changes in predictive content. Section 5 describes and provides a discussion of four empirical research questions that shed light on important issues in macroeconomic forecasting. Section 6 presents results of a meta-analysis to enhance our contribution to the practice of macroeconomic forecasting.

2 Forecasting Methods

In this section, we describe the forecasting methods compared in this paper. In contrast to the recent evaluations on forecasting output and inflation by, respectively, Chauvet and Potter (2013) and Faust and Wright (2013) we use the same set of forecasting model classes for predicting output growth and inflation. The advantage of this approach is that we can evaluate whether we need different forecasting models for output and inflation. The disadvantage is that we do not evaluate forecasting methods that were designed for some specific features of each variables, such as the UCSV models for inflation (Stock and Watson, 2007) and Markov-Switching models for output (Chauvet, 1998). Another important feature of our forecasting exercise is that we consider both point and density forecasts. The advantage of considering both point and density forecasts is that we can assess whether the choice of loss function has an impact on model rankings. Density forecasting evaluation provides us with insights on the ability of forecasting models to measure forecasting uncertainty.

In the remaining of this section we describe how we compute density forecasts of three reduced-form forecasting models: Factor models, Bayesian VAR models and MIDAS models. We also describe how we obtain density forecasts using a structural DSGE model, and simple univariate models.

In the text bellow, we use the following notation. Q_t for $t = 1, ..., T$ denotes the raw data; and $q_t = \log(Q_t)$ denotes the time series in log-levels. The variable in first differences is $\Delta q_t = 100 * (q_t - q_{t-1})$. A forecast horizon is h, and the maximum forecast horizon is $h_{\max}$.

2.1 Univariate Models

We compute density forecasts from two univariate models: a random walk model and an autoregressive (AR) model.

For the random walk, we compute point forecasts for quarterly growth rates as $\Delta\hat{q}_{T+h} = \sum_{i=0}^{3} \Delta q_{T-h-i}$, where T is the forecasting origin. We compute density forecasts by first computing the residuals for each h as

$$\hat{\varepsilon}_{h,t} = \Delta q_t - \sum_{i=0}^{3} \Delta q_{t-h-i} \text{for } t = h + 5, ..., T.$$

Then we obtain one density draw as $\Delta\hat{q}_{T+h} + v_{h,t}^{(i)}$ where $v_{h,t}^{(i)}$ is drawn by Wild bootstrap from the residuals $\hat{\varepsilon}_{h,t}$.

For the AR(p), we select the autoregressive order using the Schwarz (SIC) information criterion and assuming maximum order of 4. We compute the predictive density by bootstrap as in Clements and Taylor (2001). First, we get a full bootstrapped time series $\Delta q_{p+1}^*, ..., \Delta q_T^*$ by using the OLS estimates, initial values $\Delta q_1, ..., \Delta q_p$ and a $T - p$ bootstrapped time series from the residuals. Using the bootstrapped time series, we estimate an AR(p) model with the same autoregressive order as the original model. Then we compute forecasts by iteration for $h = 1, .., h_{\max}$ including a bootstrap draw from the residuals for each horizon. This bootstrap procedure will deliver sequential draws as $\Delta\hat{q}_{T+1}^{(i)}, ..., \Delta\hat{q}_{T+h_{\max}}^{(i)}$ for each time we reestimate the model on a new bootstrapped sample.

2.2 Factor Models

We forecast with factors using the following FADL(p,q) equation for each horizon h:

$$\Delta q_t = \beta_0 + \sum_{i=0}^{p-1} \beta_{i+1} \Delta q_{t-h-i} + \sum_{j=1}^{r} \sum_{i=0}^{q-1} \gamma_{j,i+1} f_{j,t-h-i} + \varepsilon_t, \tag{1}$$

where r counts the number of factors f.

Factors are estimated by principal components applied to either a medium (around 14 variables) or large (around 100 variables) dataset of predictors of q_t. Before the factor estimation, we decide on whether transforming raw data to log-levels as described in the "log vs level" column in Tables D2 and D2 in the online appendix. Then we apply ADF unit root tests to define the order of differentiation of each variables. Principal components is then applied to standardized data to compute the factors. We follow Groen and Kapetanios (2013) to choose the number of factors. We first choose the autoregressive order p in a univariate regression using the SIC, then we set $q = 1$ to choose the number factors using Groen and Kapetanios (2013) modified SIC assuming a maximum number of factors of 4. We have also tried to jointly choose r and q using the modified SIC, and normally $q = 1$ is the choice indicated, and even when q should be larger, the impact on average forecasting performance is negligible.

We compute density forecasts from the FADL model by fixed regressor bootstrap. We choose this specific approach because it takes into account both parameter and forecasting uncertainties when computing density forecasts, and because we will apply a similar approach, based on Aastveit, Foroni and Ravazzolo (2014), to compute density forecasts with MIDAS models. This implies that we fix the variables in the right-hand side (RHS) of the regression to their data values, and use bootstrapped values from the residuals to get a full bootstrapped time series $\Delta q^*_{p+1}, ..., \Delta q^*_T$ for the left-hand side (LHS).[6] Then we re-estimate the ADL regression using the bootstrapped LHS values and the fixed RHS values. Using bootstrapped coefficients, we compute a forecast draw $\Delta \hat{q}^{(i)}_{T+h}$,conditional on observed values for $..., \Delta q_{T-1}, \Delta q_T$, and using a bootstrap draw from the reestimated regression residuals. Note that this bootstrapping procedure will deliver the density for one specific forecasting horizon. Our factor modelling approach requires the estimation of a forecasting model for each horizon.

When predicting annual inflation, that is, $\Delta q^{(4)}_t = 100(q_t - q_{t-4})$, we use $\Delta q^{(4)}_t$ as the left-hand side variable, but we the same regressors as in equation (1).

2.3 MIDAS Models

The economic predictors in our dataset are sampled monthly. The factor approach described above requires the aggregation of monthly data into quarters. We directly exploit monthly information employing an ADL-MIDAS model. The model is written as:

$$\Delta q_t = \beta_0 + \sum_{i=0}^{p-1} \beta_{i+1} \Delta q_{t-h-i} + \gamma \sum_{i=0}^{qm-1} w(\theta, i) x_{t-mh-i+l} + \varepsilon_t,$$

[6] As a consequence, this approach does not take into account the uncertainty on the estimation of the factors, but only on the β_s and γ_s.

where m is the difference in sampling frequency between q_t and x_t, and $w(\theta, i)$ are the weights for each high frequency lag, which are a function of the parameters θ. In our applications $m = 3$ since x_t is sampled monthly while q_t is sampled quarterly. The autoregressive order in quarters is denoted by q, and qm is the autoregressive order in months such that lags of x are counted in months. The number of lead months is represented by l (named as in Andreou, Ghysels and Kourtellos (2013), but first employed for macroeconomic forecasting by Clements and Galvão (2008)). The intuition on the use of leads is that forecasts for current and future quarters are computed conditional on monthly observations of economic indicators during the current quarter. In the forecasting exercise, we set $l = 2$ for all h. This implies that we are considering typical nowcasting horizons if $h = 1$. This utilisation of monthly data is the main advantage of the MIDAS approach for macroeconomic forecasting (Clements and Galvão, 2008; Kuzin, Marcellino and Schumacher, 2013; Andreou, Ghysels and Kourtellos, 2013).

To measure the impact of the high frequency x_t on the low frequency q_t we first apply the weights $w(\theta, i)$ to all monthly lags, then we multiply by an intercept γ, which is identified because the weights sum up to one. We use the beta function to obtain the weights, that is,

$$\begin{aligned} w\left(\theta; i\right) &= \frac{f(\theta; i)}{\sum_{j=1}^{K} f(\theta; j)} \\ f(\theta; i) &= \frac{(k)^{\theta_1 - 1}(1-k)^{\theta_2 - 1}\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}; \;\; k = i/qm. \end{aligned}$$

The two parameters in θ are jointly estimated with the other parameters by nonlinear least squares. Note that, as in the case of the factor approach, we need to estimate a MIDAS regression for each forecasting horizon.

We compute density forecasts by fixed regressor bootstrapped as in Aastveit, Foroni and Ravazzolo (2014) and as described in section 2.2. Our application of the fixed regressor bootstrap to MIDAS models implies that we also fix θ, that is, take $\theta = \hat{\theta}$ from the estimation with observed data, and we obtain different values of β_i and γ for each bootstrapped sample. This has a large beneficial impact on our computational burden. Our density computation strategy is still able to capture the impact of parameter uncertainty on a set of parameters while computing forecasts. Note that, as in the case of factor models, the last step to compute $\Delta\hat{q}_{T+h}^{(i)}$ requires also a draw from the residuals of the re-estimated MIDAS regression.

We consider two different types of MIDAS specifications that are able to deal with large datasets. The first one assumes that x is an individual predictor. Because we plan to exploit sizeable datasets, we estimate a single regressor MIDAS models for each predictor, then we combine their predictive densities using equal weights. We call this model the combination MIDAS (C-MIDAS) model. In this specification, we decide beforehand whether we will be using log, log-levels or quarterly differences for each one of the indicators when using our medium dataset. Our choice of data transformation is indicated in Tables D2 and D3 in the online appendix.

The second specification estimates factors with monthly data by principal components applying the data transformation based on unit root tests described for FADL models. Then we set the number of factors to one in the case of medium datasets (14 variables) and to two in the case of large datasets following Andreou, Ghysels and Kourtellos (2013). We call this specification the F-MIDAS model, and the regressors x_t are factors estimated in a previous step by principal components

2.4 BVAR Models

Our BVAR approach is the benchmark model of Carriero, Clark and Marcellino (2013), who provide a summary the literature on the application of BVARs for forecasting. Define the vector: $y_t = (q_{1t}, q_{2t}, ..., q_{Nt})'$, then a VAR(p) is:

$$\begin{aligned} y_t &= A_0 + A_1 y_{t-1} + ... + A_p y_{t-p} + \varepsilon_t \qquad (2) \\ \varepsilon_t &\sim N(0, \Sigma) \end{aligned}$$

for $t = p+1, ..., T$.

We employ an Normal-Inverse/Wishart prior set up:

$$\begin{aligned} \alpha|\Sigma &\sim N(\alpha_0, \Sigma \otimes \Omega_0) \\ \Sigma &\sim IW(S_0, v_0), \end{aligned}$$

where $\alpha = vec([A_c, A_1, ..., A_p]')$, so the posterior distributions are

$$\begin{aligned} \alpha|\Sigma, data &\sim N(\overline{\alpha}, \Sigma \otimes \overline{\Omega}) \\ \Sigma|data &\sim IW(\bar{S}, \bar{v}). \end{aligned}$$

Carriero, Clark and Marcellino (2013) describe the close form solutions for the posterior means and variances. They are a combination of OLS estimates and the prior means weighted by the prior variances. The prior mean and variance assumptions follow a Minenessota-style prior:

$$\alpha_0 = E[A_k^{(ij)}] = \begin{cases} 1 \text{ if } i = j,\ k = 1 \\ 0 \text{ otherwise} \end{cases}$$

$$\Omega_0 = var[A_k^{(ij)}] = \begin{cases} \left(\frac{\lambda_1 \lambda_2}{k} \frac{\sigma_i}{\sigma_j}\right)^2 ,k = 1, ..., p \\ (\lambda_0 \sigma_i\,)^2 \text{ if } k = 0 \end{cases}$$

The above prior mean is appropriate for a VAR in levels. We also consider VAR specifications for growth rates defined as $\Delta y_t = y_t - y_{t-1}$. For growth rates, we set $E[A_k^{(ij)}] = 0$ for all k, i, j. The values for the sigmas are computed using univariate autoregressive models. λ_1 is the overall shrinkage parameter.

We also add in the case of VAR in levels the sum of the coefficients prior. The data is augmented with dummy observations. These are artificial observations added to the top of the data matrices y and x, when x includes lags of $y.\bar{y}_0$ is $N \times 1$ vector of the average of the first p observations:

$$\underset{N\times N}{y_d} = diag\left(\frac{\bar{y}_0}{\lambda_3}\right); \quad \underset{N\times(1+Np)}{x_d} = [0, y_d, ..., y_d]$$

If λ_3 goes to infinity, model has no cointegration and the system will be equivalent to be expressed in terms of differenced data. It means that there is unit root but no cointegration

The 'dummy initial observation' prior is also included. This was proposed by Sims (1993) and avoids giving an unreasonably high explanatory power to the initial conditions, a pathology which is typical in nearly nonstationary models (Sims, 2000).

$$\underset{1\times N}{y_{dd}} = \left(\frac{\bar{y}_0'}{\lambda_4}\right); \quad \underset{N\times(1+Np)}{x_{dd}} = [\frac{1}{\lambda_4}, y_{dd}, ..., y_{dd}]$$

This prior is consistent with cointegration if λ_4 is large. These last two priors together tend to improve forecasts when dealing with data in levels.

The prior scale matrix S_0 is assumed to be diagonal with diagonal elements given by

$$S_0^{(ii)} = (v0 - N - 1)\sigma_i^2 \text{ and } v_0 = N + 2.$$

All lambdas are equal to 1, except the one of the overall prior tightness λ_1 which is selected to maximise the marginal likelihood.

$$\lambda_1 = \arg\max_{\lambda_1} \ln(p(Y)),$$

where $p(Y)$ is computed in close form as in Carriero, Clark and Marcellino (2013). The grid has 15 elements [0.01, 0.025, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.75, 1, 2, 5]. In an out-of-sample forecasting exercise, we compute λ_1 at each time we re-estimate the model with a longer sample period.

Forecasts are computed by simulation. We use posterior draws of α and Σ to obtain a implied path for $\hat{y}_{T+1}, ..., \hat{y}_{T+h}$. Assume that $\mathbf{A} = [A_c, A_1, ..., A_p]'$ that is a $N \times Np + 1$ matrix , then we obtain a draw j for all autoregressive coefficients using:

$$(\mathbf{A}^{(j)}) = (\overline{\mathbf{A}}) + chol(\overline{\Omega}^{(j)}) * V^{(j)} * chol(\mathbf{\Sigma}^{(j)})',$$

where $V^{(j)}$ is $(Np+1) \times N$ matrix obtained from a standard normal distribution. Then for a draw of $\mathbf{A}^{(j)}$ and $\mathbf{\Sigma}^{(j)}$, we draw a sequence of h draws from the $N(0, \mathbf{\Sigma}^{(j)})$ to compute by iteration a sequence of forecasts $\hat{y}_{T+1}, ..., \hat{y}_{T+h}$ for model (2). We use a total 5000 draws, and we only accept draws if the maximum eigenvalue of the VAR companion matrix based on $\mathbf{A}^{(j)}$ is smaller than 1. We split the procedures such that we use a few number of draws of $\mathbf{A}^{(j)}$ and $\mathbf{\Sigma}^{(j)}$, and then for each parameter draw, we generate many sequences of forecasts. This strategy reduces computational time in the case that we only find stationary draws of $\mathbf{A}^{(j)}$ very infrequently as it is the case of models in levels with a large number of variables employing European data. The point forecast is the median over all draws for each horizon.

We consider specifications in levels, and we call L-BVAR, and in differences, called D-BVAR. We set $p = 4$ for specifications in first differences, and $p = 1$ for specifications in level. When the target forecasting variable is the quarterly growth rate, we transform accordingly the forecasts for the model in levels. When the target forecasting variable is the annual growth rate, we transform forecasts of both specifications accordingly. For the model in growth rates, we use that $\Delta q_t^{(4)} = 100(q_t - q_{t-4}) \approx \sum_{i=0}^{3} \Delta q_{t-i}$.

2.5 DSGE Models

The literature provides evidence of accuracy of the medium-sized Smets and Wouters (2007) model (Edge and Gurkaynak, 2011; Del Negro and Schorftheide, 2013; Woulters, 2012). We employ the Smets-Wouters DSGE model with seven observables, including output and inflation as our structural model. We use the specification in Smets and Wouters (2007) and Herbst and Schorftheide (2012), which assume a deterministic trend to productivity.

We use the priors as in Smets and Wouters (2007) and Herbst and Schorftheide (2012). The posterior distribution of the structural parameters is obtained by the Random Walk Metropolis Algorithm described in Del Negro and Schorftheide (2011), and we calibrate the spread parameter such that the acceptance rate is in the 20-40% range for each country dataset. We use 5000 draws from the posterior distribution of the parameters to compute the predictive density. For each parameter draw, we also draw from the normal distribution of the disturbances (structural shocks) to get a sequence of forecasts from $h = 1,...,h_{\max}$ for each observed variable.

We compute forecasts with DSGE models for only three countries in our dataset: US, UK and Euro Area. The reason is that the assumption in the model that the central bank that sets interest rates based on a Taylor rule, which depends on domestic inflation, is not adequate to countries which are part of the Euro Area. We also choose not apply to Japan, again because the Taylor rule may be a very poor approximation of Bank of Japan monetary policy in the last 20 years. To apply the model to Euro area data, we add an equation linking employment to hours such that we can use the employment time series instead of hours, following the modification proposed by Christoffel, Coenen and Warne (2008).

3 Data Description

We employ data from seven developed economies: US, UK, Euro Area, Germany, France, Italy and Japan. Our target variables are the quarterly change in log real GDP and the quarterly change in seasonally-adjusted log CPI with data sources described in Table D1 in the online appendix. Seasonally-adjusted CPI data is not available for European countries and Japan. As a consequence, we seasonally adjusted data using the X12 filter. Faust and Wright (2013) also employ quarterly changes in prices to measure inflation, but Bańbura, Giannone and Reichlin (2010) and Stock and Watson (2008) use the accumulated change in prices up to horizon h. In addition to quarterly change in prices, we also measure inflation as the annual change in prices, that is, $\pi_t = 100(\log(P_t) - \log(P_{t-4}))$. This additional target variable is the headline inflation in countries in which seasonally-adjusted CPI values are not published by the national statistical office.

For each country, we build a medium and a large dataset of economic indicators sampled monthly. They are described in detail in Tables D2 and D3 of the online appendix. When quarterly data are required, we use the average over quarter for factor models, so F-MIDAS nest FADL models, and the end of the quarter value for the BVAR as it is popular in the BVAR literature. When possible, we follow the series included in Kuzin, Marcellino and Schumacher (2013) datasets. The medium dataset includes 13-14 variables per country. They are a mix of measures of economic activity, including survey data, prices and financial variables. Similar set of variables have been employed by Carriero, Clark and Marcellino (2013). These datasets include oil prices as a common variable.

The number of variables included in the large dataset varies across countries due to data availability. The large dataset includes also all variables in the medium dataset. In the case of the US, we have 155 variables. Because of the international transmission of business cycles shocks, we include some key US variables in the large dataset of the 6 remaining economies, including financial variables such as equity prices and Treasury bond rates. Datasets for Germany and Italy also have more than 100 variables, while the Euro Area dataset has 81 variables. Smaller large-sized datasets are available for Japan, France and the UK, and they include around 60 variables. We provide the description of all variables including their datastream code in the Table D3 in the online appendix. Some variables were seasonally adjusted by the X12 filter before estimation, and they have SA indicated in Table D3.

DSGE models are estimated using quarterly changes in output per capita. They also use inflation measured by the GDP deflator. As consequence, when evaluating forecasts of DSGE models, we change the target variable to growth in output per capita and quarterly GDP deflator inflation. We reestimate forecasting models for these modified target variables for a subset of our reduced-form models to be able to compare predictions of structural and reduced-form models. Table D4 in the online appendix describes the variables employed in the DSGE estimation, including their required transformation.

The last observation employed in our forecasting exercise is 2013Q3. For US, Japan and UK, we use data from 1975Q1 (with exception of UK CPI inflation which is only available from 1980Q1). Observations begin in 1983Q1 for France, 1990Q1 for Italy, 1991Q1 for Germany and 1998Q2 for the Euro Area. When estimating DSGE models, we are able to use a long dataset for the Euro Area (Area-Wide Database), but for all other models we use the shorter dataset built at the monthly frequency. Data for DSGE estimation is from 1984Q1 for the US, UK and Euro Area.

4 Evaluation Design

Our first forecast origin is 1993Q1 for US, UK, Japan and France; for Germany and Italy is 1998Q1, and for the Euro Area is 2003Q1. We set the maximum forecast horizon to 8, so we are able to compute measures of forecast accuracy for forecasts up to 2011Q3, that is, we have 75 observations in our out-of-sample period for US, UK, Japan and France; 55 observations for Germany and Italy, and 35 observations for the Euro Area. For some of our results, we split the out-of-sample period in windows of 5 years (20 observations) based on the forecast origin date to verify whether the relative forecasting performance varies over the out-of-sample period. The literature provides evidence that predictive ability may change over time (Giacomini and Rossi, 2010). In addition, changes in the underlying structure of the economy and data characteristics may affect the relative forecasting performance of models. The disadvantage of short subperiods is that it is harder to find statistically significant differences in forecasting performance because of the small number of observations. As a consequence, we will also compute test statistics for the full sample period.

We compute forecasts from models estimated with expanding samples over the out-of-sample period, that is, at each forecast origin we re-estimate each model and we use all observations available up to the forecasting origin.

We use two measures of forecasting performance. Root Mean Squared Forecast Errors (RMSFE) measure the accuracy of point forecasts. The log predictive score measures the accuracy of density forecasts. The advantage of using log scores to compare density forecasts is that the maximisation of the logscore is equivalent to minimise the Kullback-Leibker distance between the model and the true density. To compute log scores, we first fit a Gaussian kernel density to the 5000 draws to a grid of values from -15 to 15. Then we compute the log score by finding the probability at the outturn.

We use the Diebold and Mariano (1995) t-statistic to test if a model is statistically more accurate than the benchmark model. We compute the t-statistic such that significant negative values imply that the model is more accurate than the benchmark. The variance is computed with the Newey-West estimator with maximum order increasing with the horizon, and we use critical values from the normal distribution. In many instances, we employ one-sided decision rules as suggested by Clark and McCracken (2013) to obtain a reasonable power level in small samples.

In addition to the univariate measure of point forecasting accuracy (RMSFE), we also consider a multivariate measure of forecasting accuracy. We employ the generalised forecasting error squared measure (GFESM) proposed by Clements and Hendry (1993).

We also evaluate the calibration of the density forecasts. If density forecasts approximate well the true data density, probability integral transforms (PITs) should be uniform, implying that the predictive density is well calibrated. We use the test proposed by Berkowitz (2001) to assess uniformity while imposing no restriction on the serial correlation of PITs over time as in Clements (2004). This implies that we can evaluate the calibration of density forecasts at all horizons.

Table 1 provides a short description of all forecasting models we employ in this evaluation. Similarly to Bańbura, Giannone and Reichlin (2010), we consider BVAR models of three sizes: small, medium and large. We use medium and large datasets for the FADL and MIDAS models, but our only small model is the BVAR. The model has only three variables: real GDP, CPI, and the short-term interest rate.

5 Forecasting Evaluation

We provide acronyms for all forecasting models included in this evaluation in Table 1. They comprise 13 reduced-form models, including an univariate model (AR), and a structural model (DSGE). We will use this set of models to answer four empirical research questions that we will discuss in detail in the remaining of this section. Our first question is to evaluate whether multivariate models, which exploit the predictive content of a set of indicators, are able to outperform simple univariate models. We also evaluate a set of circumstances that supports the choice of a specific class of models (FADL, BVAR, MIDAS and DSGE). The third question requires observing the relative performance of structural (medium-sized DSGE models) and reduced-form models. The fourth question concerns the impact of the information set available for forecasting: is it worth to employ large datasets rather than medium datasets? We answer this question by looking at measures of accuracy of point and density forecasts and at the calibration of density forecasts.

The remaining of this section is divided in subsections following the research question ordering just described. To answer these questions we make use of a set of Tables and Figures which are described when they are first mentioned.

5.1 The Predictive Content of Economic Indicators

We rank the reduced-form models in Table 1 for each target variable (output growth, quarterly inflation, annual inflation) by their performance, measured by RMSFE, logscore and GFESM, for each country, horizon and out-of-sample subperiod (93-97; 98-03; 04-07; 08-11). Then for each ranking, we make note of the models in the top 3. Table 2 presents the frequency that a model in the first column appears on the top 3 divided by the number of cases[7] for horizons $h = 1, 4$ and 8. Table 2A presents results for ranks based on MSFE, and Table 2B based on logscore. Table 2C shows ranking results based on the generalised forecasting matrix (GFESM), that is, a multivariate accuracy measure for point forecasts of output growth and quarterly inflation (Clements and Hendry, 1993).

Table 2 provides a first indication on the relative importance of multivariate models in comparison with the univariate model (AR). The results in Table 2C clearly show that the rank of the AR model relative to multivariate models improves with the horizon, suggesting that the predictive content of economic indicators decreases with the horizon. A similar result, but less clear, is found ranking by RMSFEs and logscores in Tables 2A and 2B. The ranking based on density forecasting precision, that is, on the logscore, tends to favour the AR model more than if we employ RMSFEs. Overall, the AR is not the model that most frequently appears in the top 3, providing support for predictive power of economic indicators for output growth and inflation.

The results in Table 2 have the caveat that they do not take into account if models are statistically more or less accurate than the AR model. Figures 1 and 2 present box plots of the Diebold and Mariano t-statistics computed for the full sample out-of-sample period (93-11). Negative values mean that the model is more accurate than the AR model. Using an one-sided test we would reject the null of predictability at 5% if the DM t-statistic is smaller than -1.65. The t-stats are aggregated over countries within a model class, with the name of specifications per class described in the Figure 1's notes. The t-stats distributions are presented separately for three horizons ($h = 1, 4$ and 8). Figure 1 has three plots for each horizon and for each target variable using the quadratic loss function (MSFE) to compute the t-statitics. The plots in Figure 2 instead are based on the differences in logscore.

Using both quadratic loss function for point forecasts and the logscore for density forecasts, we find predictive content for both measures of inflation for all horizons, but output growth is not predictable at the two-year horizon. Because of our large set of models and countries' datasets, these results suggest that we have to rely only on past history of output growth when predicting output growth at long horizons since it is hard to find economic indicators with long-term predictive power. Note however, we find evidence of economic indicators predictive ability at one-year horizon, in disagreement with Chauvet and Potter (2013) suggestion that we can find indicators to predict

[7] The total number of cases is 72 since we cannot rank performance for some countries for the first two subperiods due to data availability.

output growth only up to two-quarter horizons. The fact that predictive content for inflation is found at longer horizons than for output growth is compatible with the results of Mitchell, Robertson and Wright (2015) that suggest to use the R^2 of an ARMA model as guidance, while Lahiri and Sheng (2010) show that the inflation R^2 is higher than output growth at long horizons. A similar argument was made based on survey data by Patton and Timmermann (2011).

These results are robust to the use of the random walk (RW) model as a benchmark univariate model. Tables A1 and A2 in the online appendix show RMSFEs and logscores for the RW and AR models for $h = 1, 2, 4, 8$ and each country over the four five-year subperiods. The alternative multivariate model is the FADL_M model. It is clear that the AR model is more accurate univariate benchmark than the RW model for all target variables.

The first row of Table 3 summarises the evidence of Tables A1 and A2 in the online appendix: they show the percentage frequency that the factor model with a medium dataset (FADL_M) is statistically more accurate than the AR model, with frequency collected over horizons and countries. Results are presented for each subperiod and also for the full sample. The left panel employs MSFEs as measure of accuracy while the right panel compares logscores. They allow us to check how the predictive content of economic variables for each target variable has changed over time as suggested by D'Agostino and Surico (2012) when forecasting inflation. The horizontal panel present results for each one of the three forecasting targets. For both point and density forecasts, gains over the AR increase over time for output growth, and are at their highest in the last turbulent period (08-11). In contrast, for inflation, these gains decrease over time, and are really small in the last period.

In summary, we find evidence that economic indicators have predictive content for output growth at horizons up to one-year and during the most recent period, while we find predictive content for inflation at all horizons, although this predictive power has decreased in the most recent period.

5.2 Choosing a class of models

The results in Table 2 and Figures 1 and 2 also help us to indicate which model class (FADL, MIDAS, BVAR and DSGE) performs best depending on the forecasting horizon and target variable.

The MIDAS specifications have information advantage in comparison with other approaches since we use a two-month lead. Table 2A shows that this advantage produces more accurate forecasts of output at $h = 1$, appearing in 20% of top 3 slots available (C-MIDAS_M) when employing the MSFE, confirming the results by Clements and Galvão (2008) and Andreou, Ghysels and Kourtellos (2013) with US data. The results in Table 2B, however, suggest that MIDAS density forecasts are not as frequent in the top as point forecasts of output growth. When evaluating density forecasts, MIDAS models outperform other models for nowcasting quarterly inflation. At longer horizons ($h = 4, 8$) Bayesian VAR specifications or factor models are a better choice. This is clearer if we look at the results based on the GFESM in Table 2C. Best choices are C-MIDAS_M for $h = 1$, but BVAR_M and BVAR_S for longer horizons. Our ranking results are muddled at longer horizons since we do not find models that appear more 20% of the times in the top 3. [keep the phrase that follows?] In agreement with Patton (2015), model rankings change with the loss function, suggesting that our set of models are in general mispecified.

We can use the median t-statistic in Figures 1 and 2 to evaluate how each class of model performs on average across specifications and countries for each horizon. As a model class, DSGE models perform well in predicting inflation at horizons 4 and 8 for both point and density forecasts. These results, however, have to be taken with caution because we compute box-plots over three values for DSGE models but over 42 values for BVARs.

Among reduced-from models, MIDAS models do better at $h = 1$ for output growth and quarterly inflation, but fail to perform when predicting annual inflation. A possible explanation for this appalling performance of MIDAS models at the shortest horizon is that annual inflation, in contrast to quarterly inflation and output growth, is very persistent so an univariate regression has a high R^2. When attempting to estimate the MIDAS weighting function within a regression with autoregressive terms, we find problems in identifying weight function parameters when the slope parameter is small and near zero.

For one-year and two-years ahead forecasts, BVARs do better for predicting output growth and inflation. Results in favour of the BVAR in comparison with Factor and MIDAS classes are clearer when evaluating density forecasts in Figure 2.

In summary, we provide evidence that MIDAS models are the adequate choice at the nowcast horizon ($h = 1$) for predicting quarterly output growth and somewhat quarterly inflation, but at longer horizons the BVAR model class would be preferable, in particularly if one is interested in density forecasts.

5.3 Structural vs reduced-form models

The results in Figures 1 and 2 suggest that DSGE models are able to significantly improve AR forecasts of quarterly inflation at $h = 4, 8$. Note that DSGE models are estimated with output growth per capita and GDP deflator inflation instead of CPI inflation, so we re-estimate AR model with the required series when performing this comparison.

Table 3 presents additional information to help us to evaluate the relative DSGE performance. The table shows results for each one of the five-year subperiods separately and also for the full sample. The results are aggregated across countries and horizons, and we have results for a smaller subset of countries for the first two subperiods, while presenting results for only US, UK and Euro area. The entries are the frequency that the null of equal accuracy is rejected in favour of the model under the alternative. These frequency values are computed using tests for statistical differences in mean squared errors and in logscores. The last rows of the output growth and quarterly inflation panels show the number of times that the FADL_M model is statistically less accurate than the DSGE model. We can see that this situation is quite unusual when predicting output growth but it is more frequent if predicting inflation. In both cases, the DSGE model performs better in the earlier period (1993-2002) than in the later period (2003-2011), confirming the literature that supports DSGE forecasts during the Great Moderation period (1985-2007) (Del Negro and Schorftheide, 2013).

These results are supported by detailed Tables by country and forecasting horizon in the online appendix. Table A3 shows the relative performance of the DSGE model against the AR and the FADL_M using RMSFEs and Table A4 shows results with the logscore. They indicate that DSGE gains for forecasting inflation are mainly for the US and the UK, with disappointing results for the Euro area in agreement with Smets, Warne and Wouters (2014).

In summary, we provide evidence that DSGE models can deliver superior long horizon forecasts of US and UK inflation.

5.4 Dataset size

Table 3 presents the frequency that the test of equal accuracy with the FADL_M under the null is rejected against many alternative specifications. We chose the FADL_M as benchmark instead of the AR because in this section we focus on the evaluation of dataset sizes.

The first block in each panel of the table has the BVAR with only three variables (output, prices and short-rate) under the alternative. There is limited evidence that the small BVAR improves forecasts; gains from the BVAR with small dataset are more likely to happen if predicting annual inflation or during the earlier subperiods when sample sizes are shorter.

The second block looks at alternative specifications using the same medium-sized dataset. Detailed results on this block are available in Tables A5 and A6 in the online appendix. The MIDAS specifications significantly improve forecasts in particularly in the early periods. For annual inflation, the gains are mainly for long horizon forecasts, while gains are found at all horizons for the other two targets. Forecast gains from employing alternative medium-sized specifications are most likely to be detected with the US, France, UK, and Japan datasets.

The third block displays results for alternative models estimated with large-sized datasets. Detailed results on this block are available in Tables A7 and A8 in the online appendix. We find limited evidence that large datasets improve forecasts. The evidence available is spread out across countries, horizons and target variables, but it is stronger in the last two periods when sample sizes employed in the estimation are longer. The model that seems to perform better with the large datasets is the combination MIDAS model (C-MIDAS_L), supporting the claims that it might be worth to estimate one forecasting model for each indicator and then combine forecasts than accommodate a large number of indicators within a unique forecasting model (as the BVARs and FADL_M) (Clements and Galvão, 2006; Rossi and Skhposyan, 2014). This holds for both point and density forecasts.

Figure 3 provides additional evidence on the comparison between medium and large datasets. We present Diebold and Mariano t-statistics with the model with a medium dataset under the null and the model with the large dataset under the alternative. We present results for the following models: FADL, F-MIDAS, C-MIDAS, L-BVAR and D-BVAR. The box plots are computed for t-statistics varying over horizons ($h = 1, ..., 8$) and countries, and are computed over the full out-of-sample period. Negative values imply that the model with a large dataset is more accurate than those with a medium data set. Using a two-sided 5% test, statistical differences are found when the absolute value of the t-stat is larger than 1.96.

For output growth, we are more likely to find statistical differences in favour of the large dataset with the FADL and the F-MIDAS, but the C-MIDAS is a better alternative for quarterly inflation, and the FADL for annual inflation. In general, the t-statistics are between -1.96 and 1.96, that is, models with large and medium datasets deliver statistically similar point and density forecasting performances.

In summary, we cannot provide overwhelming evidence that large datasets improve forecasts over medium datasets, although this might occur for some countries and horizons. Significant gains from the use of large datasets are more likely to occur with the inclusion of monthly factors in MIDAS models when predicting output growth, and using combination MIDAS models when predicting inflation. Gains are also more likely to occur during the most recent period (1998-2011).

5.4.1 Dataset size and the Calibration of Density Forecasts.

Our previous assessments make use of logscores to measure the relative forecasting accuracy of density forecasts. An issue of this approach is that although model A may be more accurate than model B, both models might be a bad approximation of the underlying true data density. Even if two models approximate the true data density, it might be that one of them has better sharpness as measured by logscores (Mitchell and Wallis, 2011), as reported in the previous subsections. In this subsection, we look at the results of the Berkowitz test as described in section 4. Failure to reject the null implies that density forecasts are well calibrated.

Table 4 presents the p-values of the chi-squared Berkowitz (2001) test for uniformity for all models in Table 1, presented for each country and for horizons $h = 1, 4$ and 8. Entries shaded in green indicate that the null of uniformity is not rejected at the 10% level. Table 4A presents results for output growth, and Tables 4B and 4C for, respectivelly, quarterly and annual inflation. The test statistics were computed using PITs for the full out-of-sample period (93-11).

In general, models with small datasets (AR and BVAR_S models) are able to provide well calibrated densities. However, we can point to some cases in which a large set of regressors improved the calibration of density forecasts: when predicting the US and UK output growth and quarterly inflation. In the case of output growth, models with large datasets are more useful, but models with medium datasets perform better for inflation. The online appendix shows PIT's histograms in five bins for each country in Figures A1 to A4. We present results for one representative model for each dataset size (AR, L-BVAR_S, C-MIDAS_M and F-MIDAS_L). Figures A1 and A2 are based on forecasts of output growth at horizons 1 and 4 and Figures A3 and A4 are for forecasts of quarterly inflation at horizons 1 and 4. They have an inverted U-shape for one-quarter and one-year-ahead forecasts of US output growth. This implies that density forecasts are too wide. Clark (2011) shows how inclusion of stochastic volatility in the disturbances of small VAR models is able to solve this problem, while Diebold, Schorftheide and Shin (2015) provide similar results with DSGE models. Here we provide new results: by enlarging the number of predictors from 1 to 155 (see details on the data appendix), we are also able to provide well calibrated density forecasts of US output growth and inflation.

In summary, by employing a large dataset, we are able to deliver well-calibrated density forecasts of US and UK output growth, and US inflation, improving over models with small datasets. In general, small models density forecasts are well-calibrated for the remaining countries (Euro area, Germany, France, Italy and Japan).

5.5 Additional results

The four research questions addressed so far are crucial aspects for the macroeconomic forecasting literature. However the results compiled in the Tables and Figures that are included in this paper may also help us to provide additional guidance for practitioners.

Table 2 and Figures 1 and 2 suggest that it is not advisable to use the same forecasting model for both output growth and inflation, supporting the development of different forecasting models for each target variable as described in Chauvet and Potter (2013) and Faust and Wright (2013). Even for the horizons for which BVARs are the chosen model class ($h = 4, 8$), the adequate specification may differ across variables. This result may be also used to support the use of reduced-form models instead of structural models when forecasting. While we provide evidence in Tables 3 and A3 and A4 that DSGE models are accurate inflation forecasters, a similar performance for output growth is not recorded.

Tables 2 and 3 suggest that we should rather combine forecasts of single indicator models (C-MIDAS) than use factors extracted for the same set of regressors (F-MIDAS) when computing point forecasts. This result is less clear-cut when looking at the accuracy of density forecasts. Note that these results are based on equal-weight combinations. The computation of weights based on past performance may improve forecasts, but we leave this issue for future research.

Table 3 provides a direct comparison between BVARs in levels and in differences. The use of BVARs in differences instead of levels provides frequent improvements in density forecasting performance during the Great Moderation period (93-02). Gains are smaller in the most recent period.

6 What explains forecasting performance? A meta analysis

In the previous section, we use our empirical forecasting exercise to answer a set of research questions with aim to contribute to the macroeconomic forecasting literature. In this section, we enhance this paper contribution to the practice of macroeconomic forecasting by using a meta analysis to measure the impact of some forecasting exercise features on forecasting performance. We measure the impact of model class, forecasting horizon, dataset size and country source on point and density forecasting performance of reduced-form statistical models. We regress forecasting performance, measured by root mean squared forecast errors and the median logscore[8], on a set of dummy variables measuring the characteristics of interest.

The dependent variable is a measure of the relative forecasting performance of a specific forecasting model to the autoregressive model when predicting one of our target variables (output growth, quarterly and annual inflation) for a specific country, horizon and forecasting origin period. The measures of forecasting performance are based on root mean squared forecast errors (RMSFE) and the median logscore (MLS) computed for a specific target variable varying across country, forecasting model, period and horizon. The measures for point and density forecasting performance are:

$$\begin{aligned} rperf1_{m,p,c,h} &= \frac{RMSFE_{AR,p,c,h}}{RMSFE_{m,p,c,h}}; \\ rperf2_{m,p,c,h} &= 1 + [(-MLS_{ar,p,c,h}) - (-MLS_{m,p,c,h})]. \end{aligned}$$

where $m = 2, ..., 13$, which are the statistical models numbered 2 up to 13 in Table 1. Each measure varies with the set of forecasting origins employed in the computation p =93Q1-97Q4, 98Q1-02Q4, 03Q1-07Q4, 08Q1-11Q3, 93Q1-11Q3; with the source country c =US, UK, EU, FR, IT, GER. JP, and the forecasting horizon $h = 1, ..., 8$. This means that the expected total number of observations for each measure (ignoring the fact we have missing values for some of the earlier forecasting origin periods) is 3360. A sample large enough to find sources of statistically significant forecasting performance improvements.

[8] We use the median logscore to minimize the impact of outliers in our analysis. Outlier values are more frequent with logscores than MSFEs.

To exploit variation in performance across countries, we use two dummy variables to split the country set into three: $D\_EU = 1$ for European countries ($c =$EU, FR, IT, GER), and $D\_JP = 1$ if $c =$JP. The benchmark countries are US and UK. To exploit the impact of the forecasting horizon, we split the set of forecast horizons into three groups by defining $D\_mh = 1$ if $h = 2, ..., 4$ and $D\_lh = 1$ if $h = 5, ..., 8$. Accordingly, differences in performance over medium and long horizons are assessed against the nowcasting ($h = 1$) benchmark. The impact of model class is assessed by using D_MIDAS=1 if $m = 4, 5, 6, 7$ and $D\_BVAR = 1$ if $m = 8, 9, 10, 11, 12, 13$ based on Table 1 description. The benchmark model class is the FADL ($m = 2, 3$). The impact of dataset size is evaluated using $D\_small = 1$ if $m = 8, 9$ and $D\_l\arg e = 1$ if $m = 3, 5, 7, 12, 13$, implying that the benchmark model has a medium dataset. Finally, the impact of the forecasting origin period is assessed by creating one dummy variable for each one of the four subperiods. As a consequence, relative performance improvements are measured with the full out-of-sample period as benchmark ($p =$93Q1-11Q3).

Table 5 presents estimates of coefficients on each set of dummy variables described above and a set of interactions between them. The table show results per performance measure (*rperf1* and *rperf2)* and target variables (output growth, quarterly inflation and annual inflation). Cases where the null hypothesis of statistical insignificance is rejected are indicated with stars, as usual.

The intercept of the regressions for $rperf1_{m,p,c,h}$ and $rperf2_{m,p,c,h}$ are estimated at a value larger than 1, implying that the FADL_M normally improves over the AR when nowcasting US and UK variables. Gains are larger for output growth and are of 4% in terms of RMSFE. Results based on the country dummies indicate that gains from employing multivariate models instead of AR models for predicting output growth are larger with Japanese data but smaller with European data. In contrast, when predicting inflation, multivariate models gains obtained with European data are as the ones with US and UK data.

The coefficients on the variables for each forecasting origin period imply changes in statistical performance over time. During the 93Q1-97Q4 period, models are relatively more accurate than the AR for inflation, but they are worse for output growth. As consequence, this earlier great moderation period favours AR for output growth but multivariate models for inflation. In contrast, during the turbulent 08Q1-11Q3 period, we find that multivariate models perform relatively better than AR models for output growth, but the single period performance is similar to the full sample for inflation.

Coefficients on the horizon dummies are all negative and are all significant when comparing long horizon with nowcasting performances. This implies that the relative performance of multivariate models to the AR model deteriorates significantly with the horizon. When comparing model classes, FADL, MIDAS and BVAR perform similarly for nowcasting, but the MIDAS model is significantly worse for annual inflation at $h = 1$ as reported in section 5.2. Note that although MIDAS models improve the performance of FADL in 3% on average when nowcasting output growth, this coefficient is not statistically significant (pvalue of 10.5%). We include a set of interactions between horizon and model class dummies. The results suggest that MIDAS models improve their forecasting performance at longer horizons when predicting annual inflation. Confirming our previous results, it is hard to define a model class that is superior; their relative forecasting performance is similar in many cases.

FADL models are significantly affected by dataset size when measuring point forecasting performance, but this is not the case when looking at density forecasts. Larger datasets make inflation forecasts significantly worse in terms of RMSFEs with losses of around 5%. Small models (L_BVAR_S and D_BVAR_S) are better than the FADL_M for computing point forecasts of inflation, but they are worse for output growth. When interacting the dataset-size with the model class dummies, we find that using a large BVAR instead of a FADL_M deteriorates significantly forecasting performance for all variables. These results confirm our previous discussion that there is no clear evidence supporting the use of a large dataset of predictors (around one-hundred) instead of a medium data set (about a dozen picked variables) for forecasting macro variables.

In summary, we find some variation in the relative forecasting performance of multivariate to AR models across countries. The relative performance of multivariate models for inflation is at its peak in the 93-97 period, but in the 2008-2011 period if predicting output growth. We find no evidence that models with large datasets improve over the performance of models with smaller datasets, and in many cases we cannot favour undoubtedly a specific model class.

7 Conclusion

The comprehensive evaluation of macroeconomic forecasting models reported in this paper contributes to the academic literature and the practice of macroeconomic forecasting. By employing datasets for seven developed economies and considering four classes of multivariate forecasting models, we provide new empirical findings, extending and enhancing evidence usually available for US data.

Our evaluation is designed to look at forecasting horizons from nowcasting up to two-years ahead. Our contribution is to consider a large set of model specifications over all these horizons so we can provide evidence that the choice of the best forecasting model class clearly varies with the forecast horizon.

An important new finding is that by increasing the number of predictors employed for forecasting output growth and quarterly inflation, we may improve the calibration of density forecasts, so the problem of too wide density forecasts due to the impact of the Great Moderation is attenuated. A possible explanation for this fact is that smaller scale models might suffer from omitted variables, and therefore the importance of time variation in volatility. that is typically found in small-sized models, might be just an artifact of this misspecification. We confirm the predictability results based on US survey data (Patton and Timmermann, 2011), who suggested that inflation is more persistent and predictable than GDP growth, by finding evidence of predictive power of economic indicators for inflation at long horizons (two-years ahead) but not for output growth. Even with our large set of countries and model specifications, we find no case in which a model with economic indicators significantly improves two-year ahead forecasts of a univariate model for output growth.

We extend results based only on Bayesian VARs (Koop, 2013) by showing that usually the application of large datasets instead of medium ones do not improve forecasts. Our contribution is to employ five different specifications from three model classes to address the interesting research question on whether it is worth to use large datasets instead of using 13-14 hand picked predictors for both point and density forecasting, and we find that indeed a medium dataset typically suffice.

Finally, our multicountry comparison provides a new dimension when comparing structural with reduced-form models in forecasting. The DSGE model specification we consider (Smets and Wouters, 2007) provides accurate forecasts not only of US but also of UK inflation when estimated with UK data.

References

Aastveit, K. A., Foroni, C. and Ravazzolo, F. (2014). Density forecasts with midas models, *Norges Bank Working Paper 10/2014* .

Andreou, E., Ghysels, E. and Kourtellos, A. (2013). Should macroeconomic forecasters look at daily financiaoutput and how?, *Journal of Business and Economic Statistics* **31**: 240–251.

Bańbura, M., Giannone, D. and Reichlin, L. (2010). Large bayesian vector autoregressions, *Journal of Applied Econometrics* **25**(1): 71–92.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management., *Journal of Business and Economic Statistics* **19**: 465–474.

Carriero, A., Clark, T. E. and Marcellino, M. (2013). Bayesian vars: Specifications choices and forecast accuracy, *Journal of Applied Econometrics* **(in press)**.

Chauvet, M. (1998). An econometric characterization of business cycle dynamcis with factor structure and regime switches., *International Economic Review* **39**: 969–96.

Chauvet, M. and Potter, S. (2013). Forecasting output, *Handbook of Economic Forecasting, volume 2A*, Elsevier, chapter 3, pp. 141–194.

Christoffel, K., Coenen, G. and Warne, A. (2008). The new area-wide model of the euro area: a micro-founded open-economy model for forecasting and policy analysis, *ECB Working Paper Series n. 944* .

Clark, T. E. (2011). Real-time density forecasts from baryesian vector autoregressions with stochastic volatility, *Journal of Business and Economic Statistics* **29**.

Clark, T. E. and McCracken, M. W. (2013). Advances in forecast evaluation, *Handbook of Economic Forecasting, volume 2B*, Elsevier, chapter 20, pp. 1107–1201.

Clements, M. P. (2004). Evaluating the bank of england density forecasts of inflation, *Economic Journal* **114**: 855–877.

Clements, M. P. and Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States, *Journal of Business and Economic Statistics* **26**: 546–554. No. 4.

Clements, M. P. and Galvão, A. B. (2006). Combining pprediction versus information in modemodel: forecasting us recession probabilities and output growth, *in* C. Milas, P. Rothman and D. van Dijk (eds), *nonlinear time series analysis of business cycles*, elsevier, pp. 55–74.

Clements, M. P. and Hendry, D. F. (1993). On the limitations of comparing mean squared forecast errors., *Journal of Forecasting* **14**: 617–37.

Clements, M. P. and Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models, *International Journal of Forecasting* **17**: 247–267.

D'Agostino, A. and Surico, P. (2012). A century of inflation forecasts, *The review of economics and statistics* **94**: 1097–1106.

Del Negro, M. and Schorftheide, F. (2011). Bayesian macroeconometrics, *The Oxford Handbook of Bayesian Econometrics* pp. 293–389.

Del Negro, M. and Schorftheide, F. (2013). DSGE model-based forecasting, *Handbook of Economic Forecasting, volume 2A*, Elsevier, chapter 2, pp. 57–140.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**: 253–263. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

Diebold, F. X., Schorftheide, F. and Shin, M. (2015). Real-time forecast evaluation of dsge models with stochastic volatility, *University of Pennsylvannia, mimeo* .

Edge, R. M. and Gurkaynak, R. S. (2011). How useful are estimated DSGE model forecasts, *Federal Reserve Board, Finance and Economics Discussion Series* **11**.

Faust, J. and Wright, J. H. (2013). Forecasting inflation, *Handbook of Economic Forecasting, volume 2A*, Elsevier, chapter 1, pp. 3–56.

Ferrara, L., Marcellino, M. and Mogliani, M. (2015). Macroeconomic forecasting during the Great Recession: the return of non-linearity?, *International Journal of Forecating* **31**: 664–679.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments, *Journal of Applied Econometrics* **25**: 595–620.

Groen, J. J. J. and Kapetanios, G. (2013). Model selection criteria for factor-augmented regressions, *Oxford Bullettin of Economics and Statistics* **75**: 37–63.

Herbst, E. and Schorftheide, F. (2012). Evaluating DSGE model forecasts of comovements, *Journal of Econometrics* **171**: 152–166.

Koop, G. (2013). Forecasting with medium and large bayesian VARs, *Journal of applied econometrics* **28**: 177–203.

Kuzin, V., Marcellino, M. and Schumacher, C. (2013). Pooling versus model selection for nowcasting with many predictors: Empirical evidence for six industrialized countries, *Journal of Applied Econometrics* **28**: 392–411.

Lahiri, K. and Sheng, X. (2010). Learning and heterogeneity in gdp and inflation forecasts., *International Journal of Forecasting* **26**: 265–292.

Mitchell, J., Robertson, D. and Wright, S. (2015). What univariate models can tell us about multivariate macroeconomic models?, *working paper* .

Mitchell, J. and Wallis, K. F. (2011). Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness, *Journal of Applied Econometrics* **26**: 1023–1040.

Patton, A. J. (2015). Evaluating and comparing possibly misspecified forecasts, *Duke University Working Paper* .

Patton, A. J. and Timmermann, A. (2011). Predictability of output growth and inflation: a multi-horizon survey approach, *Journal of Business and Economic Statistics* **29**: 397–410.

Rossi, B. and Skhposyan, T. (2014). Evaluating predictiove densities of us output growth andinflation in a large macroeconomic data set, *International Journal of Forecasting* **30**: 662–682.

Sims, C. (1993). A nine-variable probabilistic macroeconomic forecasting model, *Business Cycles, Indicators and Forecasting*, National Bureau of Economic Research, pp. 179–212.

Sims, C. (2000). Using a likelihood perspective to sharpen econometric discourse: three examples., *Journal of Econometrics* **95**(2): 443–462.

Smets, F., Warne, A. and Wouters, R. (2014). Professional forecasters and real-time forecasting with a dsge model, *International Journal of Forecasting* **30**: 981–995.

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles., *American Economic Review* **97**: 586–606.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes, *Journal of Business and Economic Statistics* **20**: 147–162.

Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices, *Journal of Economic Literature* **41**: 788–829.

Stock, J. H. and Watson, M. W. (2007). Why has U.S. Inflation Become Harder to Forecast?, *Journal of Money, Credit and Banking* **Supplement to Vol. 39**: 3–33.

Stock, J. H. and Watson, M. W. (2008). Phillips Curve Inflation Forecasts, *NBER working paper series n. 14322* .

Woulters, M. H. (2012). Evaluating point and density forecastmodel DSGE models, *Munich Personal RePEc Archive Paper* **n. 36147**.

Table 1: Model Acronyms

	Name	Description
1	AR	Autoregressive Model
2	FADL_M	Factor ADL model with medium-sized dataset
3	FADL_L	Factor ADL model with large-sized dataset
4	F-MIDAS_M	Factor MIDAS with medium-sized dataset
5	F-MIDAS_L	Factor MIDAS with large-sized dataset
6	C-MIDAS_M	Combination MIDAS with medium-sized dataset
7	C-MIDAS_L	Combination MIDAS with large-sized dataset
8	L-BVAR_S	BVAR in levels with small dataset.
9	D-BVAR_S	BVAR in differences with small dataset.
10	L-BVAR_M	BVAR in levels with medium-sized dataset.
11	D-BVAR_M	BVAR in differences with medium-sized dataset.
12	L-BVAR_L	BVAR in levels with large-sized dataset.
13	D-BVAR_L	BVAR in differences with large-sized dataset.
14	DSGE	Smets and Wouters (2007) medium-sized DSGE model.

Table 2: Frequency that the model is on the top 3: ranked per country and per period (93-97; 98-02; 03-07; 08-11).

Table 2A: Ranked by MSFE

	Output Growth			Quarterly Inflation			Annual Inflation		
Models	H=1	H=4	H=8	H=1	H=4	H=8	H=1	H=4	H=8
AR	8.3%	5.6%	8.3%	4.2%	8.3%	11.1%	8.3%	5.6%	8.3%
L-BVAR_S	4.2%	11.1%	6.9%	5.6%	12.5%	13.9%	9.7%	9.7%	12.5%
D-BVAR_S	1.4%	6.9%	6.9%	8.3%	11.1%	11.1%	16.7%	11.1%	9.7%
FADL_M	5.6%	4.2%	13.9%	5.6%	13.9%	4.2%	12.5%	8.3%	1.4%
F-MIDAS_M	4.2%	5.6%	8.3%	8.3%	4.2%	6.9%	0.0%	6.9%	6.9%
C-MIDAS_M	20.8%	6.9%	6.9%	13.9%	6.9%	11.1%	0.0%	12.5%	9.7%
L-BVAR_M	6.9%	18.1%	8.3%	12.5%	12.5%	5.6%	16.7%	12.5%	8.3%
D-BVAR_M	9.7%	6.9%	6.9%	15.3%	6.9%	6.9%	13.9%	8.3%	13.9%
FADL_L	8.3%	2.8%	4.2%	2.8%	4.2%	6.9%	6.9%	6.9%	2.8%
F-MIDAS_L	12.5%	4.2%	5.6%	6.9%	4.2%	4.2%	0.0%	1.4%	4.2%
C-MIDAS_L	2.8%	2.8%	6.9%	4.2%	6.9%	12.5%	1.4%	5.6%	13.9%
L-BVAR_L	9.7%	18.1%	8.3%	8.3%	6.9%	4.2%	8.3%	8.3%	6.9%
D-BVAR_L	5.6%	6.9%	8.3%	4.2%	1.4%	1.4%	5.6%	2.8%	1.4%

Table 2B: Ranked by Logscore.

	Output Growth			Quarterly Inflation			Annual Inflation		
Models	H=1	H=4	H=8	H=1	H=4	H=8	H=1	H=4	H=8
AR	6.9%	15.3%	12.5%	8.3%	9.7%	11.1%	13.9%	5.6%	11.1%
L-BVAR_S	0.0%	6.9%	11.1%	5.6%	8.3%	6.9%	6.9%	13.9%	12.5%
D-BVAR_S	2.8%	5.6%	4.2%	6.9%	9.7%	13.9%	13.9%	9.7%	8.3%
FADL_M	6.9%	8.3%	8.3%	8.3%	13.9%	6.9%	18.1%	6.9%	5.6%
F-MIDAS_M	5.6%	4.2%	6.9%	4.2%	9.7%	4.2%	0.0%	9.7%	9.7%
C-MIDAS_M	12.5%	2.8%	9.7%	20.8%	12.5%	9.7%	1.4%	11.1%	8.3%
L-BVAR_M	6.9%	15.3%	11.1%	2.8%	2.8%	5.6%	8.3%	9.7%	6.9%
D-BVAR_M	18.1%	4.2%	5.6%	15.3%	8.3%	6.9%	13.9%	8.3%	11.1%
FADL_L	9.7%	9.7%	6.9%	8.3%	2.8%	9.7%	11.1%	4.2%	2.8%
F-MIDAS_L	15.3%	0.0%	4.2%	4.2%	8.3%	4.2%	0.0%	4.2%	2.8%
C-MIDAS_L	1.4%	5.6%	6.9%	6.9%	9.7%	13.9%	1.4%	2.8%	8.3%
L-BVAR_L	8.3%	12.5%	6.9%	4.2%	4.2%	4.2%	6.9%	8.3%	8.3%
D-BVAR_L	5.6%	9.7%	5.6%	4.2%	0.0%	2.8%	4.2%	5.6%	4.2%

Table 2C: Ranked by GFESM (Output Growth and Quarterly Inflation)

Models	H=1	H=4	H=8
AR	4.2%	8.3%	9.7%
L-BVAR_S	4.2%	12.5%	13.9%
D-BVAR_S	2.8%	5.6%	8.3%
FADL_M	6.9%	4.2%	8.3%
F-MIDAS_M	2.8%	5.6%	5.6%
C-MIDAS_M	19.4%	6.9%	5.6%
L-BVAR_M	11.1%	16.7%	9.7%
D-BVAR_M	11.1%	5.6%	8.3%
FADL_L	6.9%	4.2%	1.4%
F-MIDAS_L	8.3%	4.2%	2.8%
C-MIDAS_L	4.2%	8.3%	9.7%
L-BVAR_L	12.5%	15.3%	9.7%
D-BVAR_L	5.6%	2.8%	6.9%

Note: The Generalised Forecasting Error Squared Measure (GFESM) is computed as the determinant of the bivariate squared forecasting error matrix. The total number of top 3 slots per horizon/target variable is 72 (4 periods X 7 countries (but not all countries for first two periods) X 3). Columns sum up to 1.

Table 3: Percentage Rejection of the null of Equal Forecast Accuracy in favor of model per period, aggregated across countries and horizons (h=1,…,8).

	Point Forecasting					Density Forecasting				
	1993Q1-1997Q4	1998Q1-2002Q4	2003Q1-2007Q4	2008Q1-2011Q3	1993Q1-2011Q3	1993Q1-1997Q4	1998Q1-2002Q4	2003Q1-2007Q4	2008Q1-2011Q3	1993Q1-2011Q3
Bench vs model	Output Growth									
AR vs FADL_M	9.38%	0.00%	7.14%	14.29%	1.79%	6.25%	2.08%	8.93%	17.86%	3.57%
FADL_M vs L-BVAR_S	25.00%	10.42%	5.36%	0.00%	8.93%	12.50%	6.25%	3.57%	8.93%	10.71%
FADL_M vs D-BVAR-S	6.25%	10.42%	0.00%	7.14%	1.79%	0.00%	12.50%	3.57%	3.57%	5.36%
FADL_M vs F-MIDAS_M	9.38%	4.17%	5.36%	14.29%	1.79%	9.38%	0.00%	3.57%	7.14%	0.00%
FADL_M vs C-MIDAS_M	15.63%	16.67%	14.29%	16.07%	26.79%	3.13%	12.50%	8.93%	10.71%	12.50%
FADL_M vs L-BVAR_M	3.13%	12.50%	8.93%	17.86%	23.21%	3.13%	6.25%	5.36%	14.29%	12.50%
FADL_M vs FADL_L	18.75%	12.50%	12.50%	1.79%	8.93%	12.50%	12.50%	7.14%	0.00%	5.36%
FADL_M vs F-MIDAS_L	0.00%	4.17%	12.50%	21.43%	10.71%	12.50%	10.42%	3.57%	10.71%	8.93%
FADL_M vs L-BVAR_L	6.25%	10.42%	8.93%	21.43%	17.86%	3.13%	4.17%	0.00%	10.71%	8.93%
FADL_M vs C-MIDAS_L	3.13%	4.17%	8.93%	12.50%	10.71%	0.00%	4.17%	1.79%	12.50%	5.36%
L-BVAR_M vs D-BVAR_M	25.00%	2.08%	0.00%	5.36%	3.57%	46.88%	4.17%	1.79%	3.57%	3.57%
L-BVAR_L vs D-BVAR_L	0.00%	2.08%	1.79%	3.57%	1.79%	21.88%	12.50%	5.36%	14.29%	5.36%
FADL_M vs DSGE	6.25%	18.75%	4.17%	8.33%	20.83%	0.00%	6.25%	0.00%	0.00%	0.00%
	Quarterly Inflation									
AR vs FADL_M	37.50%	16.67%	8.93%	3.57%	17.86%	21.88%	27.08%	12.50%	7.14%	5.36%
FADL_M vs L-BVAR_S	15.63%	20.83%	7.14%	3.57%	7.14%	6.25%	2.08%	7.14%	5.36%	12.50%
FADL_M vs D-BVAR-S	37.50%	20.83%	10.71%	3.57%	8.93%	18.75%	12.50%	8.93%	7.14%	12.50%
FADL_M vs F-MIDAS_M	21.88%	14.58%	14.29%	17.86%	8.93%	15.63%	20.83%	5.36%	25.00%	10.71%
FADL_M vs C-MIDAS_M	0.00%	4.17%	16.07%	30.36%	21.43%	12.50%	14.58%	19.64%	30.36%	26.79%
FADL_M vs L-BVAR_M	18.75%	6.25%	12.50%	5.36%	12.50%	0.00%	2.08%	1.79%	7.14%	8.93%
FADL_M vs FADL_L	6.25%	8.33%	1.79%	10.71%	3.57%	25.00%	2.08%	5.36%	8.93%	3.57%
FADL_M vs F-MIDAS_L	18.75%	8.33%	0.00%	8.93%	1.79%	25.00%	10.42%	0.00%	7.14%	3.57%
FADL_M vs L-BVAR_L	12.50%	14.58%	1.79%	3.57%	1.79%	0.00%	0.00%	3.57%	1.79%	1.79%
FADL_M vs C-MIDAS_L	21.88%	14.58%	16.07%	25.00%	26.79%	37.50%	18.75%	8.93%	23.21%	28.57%
L-BVAR_M vs D-BVAR_M	18.75%	0.00%	16.07%	21.43%	10.71%	21.88%	14.58%	28.57%	21.43%	14.29%
L-BVAR_L vs D-BVAR_L	0.00%	2.08%	12.50%	10.71%	14.29%	21.88%	14.58%	10.71%	8.93%	26.79%
FADL_M vs DSGE	31.25%	31.25%	16.67%	4.17%	25.00%	37.50%	50.00%	37.50%	12.50%	50.00%
	Annual Inflation									
AR vs FADL_M	65.63%	31.25%	5.36%	0.00%	12.50%	40.63%	18.75%	3.57%	1.79%	8.93%
FADL_M vs L-BVAR_S	3.13%	2.08%	14.29%	1.79%	17.86%	3.13%	8.33%	19.64%	5.36%	21.43%
FADL_M vs D-BVAR-S	18.75%	14.58%	16.07%	1.79%	10.71%	18.75%	12.50%	10.71%	1.79%	12.50%
FADL_M vs F-MIDAS_M	9.38%	12.50%	12.50%	8.93%	10.71%	12.50%	8.33%	0.00%	5.36%	12.50%
FADL_M vs C-MIDAS_M	0.00%	4.17%	21.43%	14.29%	3.57%	3.13%	0.00%	12.50%	5.36%	7.14%
FADL_M vs L-BVAR_M	21.88%	12.50%	28.57%	1.79%	17.86%	9.38%	10.42%	14.29%	3.57%	25.00%
FADL_M vs FADL_L	12.50%	0.00%	5.36%	17.86%	5.36%	31.25%	2.08%	1.79%	14.29%	7.14%
FADL_M vs F-MIDAS_L	12.50%	0.00%	0.00%	5.36%	0.00%	6.25%	0.00%	0.00%	0.00%	0.00%
FADL_M vs L-BVAR_L	18.75%	6.25%	10.71%	1.79%	7.14%	31.25%	0.00%	14.29%	1.79%	10.71%
FADL_M vs C-MIDAS_L	9.38%	8.33%	17.86%	16.07%	10.71%	6.25%	0.00%	8.93%	7.14%	7.14%
L-BVAR_M vs D-BVAR_M	12.50%	2.08%	7.14%	23.21%	10.71%	21.88%	8.33%	12.50%	25.00%	7.14%
L-BVAR_L vs D-BVAR_L	0.00%	4.17%	12.50%	8.93%	10.71%	6.25%	16.67%	10.71%	10.71%	14.29%

Notes: Countries: US, UK, Germany, and Japan from 1993Q1, including also Italy and Germany from 1998Q1, and the Euro Area from 2003Q1. If DSGE under the alternative only for US, UK and Euro Area. Detailed results in online appendix tables.

Table 4 P-values of the Berkowitz test for uniformity computed over full out-of-sample period (93-11)

Table 4A: Output Growth

	US			UK			Eurozone			Germany			France			Italy			Japan		
Models	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
AR	0.03	0.01	0.00	0.28	0.19	0.03	0.13	0.02	0.00	0.18	0.49	0.55	0.54	0.29	0.04	0.07	0.02	0.01	0.01	0.00	0.01
L-BVAR_S	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.15	0.06	0.74	0.84	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.10	0.14
D-BVAR_S	0.00	0.00	0.00	0.00	0.00	0.00	0.55	0.19	0.13	0.06	0.33	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.19	0.31
FADL_M	0.06	0.07	0.02	0.00	0.00	0.01	0.03	0.01	0.00	0.09	0.14	0.09	0.02	0.10	0.00	0.01	0.00	0.00	0.14	0.12	0.63
F-MIDAS_M	0.09	0.05	0.01	0.00	0.41	0.37	0.01	0.00	0.00	0.05	0.01	0.13	0.00	0.02	0.02	0.32	0.00	0.00	0.34	0.02	0.01
C-MIDAS_M	0.00	0.00	0.01	0.26	0.00	0.00	0.06	0.00	0.00	0.95	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00
L-BVAR_M	0.04	0.01	0.01	0.77	0.99	0.64	0.95	0.04	0.01	0.69	0.49	0.82	0.07	0.03	0.03	0.02	0.05	0.02	0.67	0.50	0.73
D-BVAR_M	0.08	0.00	0.00	0.04	0.01	0.00	0.05	0.10	0.17	0.10	0.28	0.84	0.20	0.09	0.02	0.03	0.02	0.00	0.01	0.01	0.00
FADL_L	0.88	0.06	0.03	0.00	0.01	0.00	0.03	0.01	0.00	0.14	0.01	0.15	0.20	0.07	0.02	0.20	0.00	0.00	0.01	0.00	0.01
F-MIDAS_L	0.09	0.37	0.13	0.99	0.22	0.36	0.10	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.13
C-MIDAS_L	0.02	0.03	0.04	0.32	0.00	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.03	0.00	0.00	0.04	0.00	0.00
L-BVAR_L	0.00	0.04	0.03	0.76	0.91	0.99	0.75	0.00	0.00	0.71	0.50	0.07	0.80	0.46	0.02	0.00	0.05	0.00	0.57	0.82	0.79
D-BVAR_L	0.00	0.01	0.01	0.28	0.36	0.35	0.02	0.05	0.11	0.01	0.07	0.00	0.00	0.01	0.05	0.00	0.06	0.01	0.00	0.01	0.00
DSGE	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00												

Table 4B: Quarterly Inflation.

	US			UK			Eurozone			Germany			France			Italy			Japan		
Models	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
AR	0.45	0.02	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.96	0.95	0.79	0.07	0.01	0.00	0.28	0.02	0.00	0.01	0.00	0.00
L-BVAR_S	0.01	0.01	0.00	0.00	0.00	0.00	0.34	0.05	0.06	0.00	0.02	0.01	0.00	0.00	0.00	0.44	0.63	0.49	0.39	0.15	0.20
D-BVAR_S	0.70	0.01	0.00	0.01	0.04	0.09	0.26	0.46	0.17	0.02	0.03	0.05	0.11	0.29	0.92	0.02	0.39	0.50	0.88	0.63	0.45
FADL_M	0.26	0.01	0.00	0.01	0.05	0.00	0.01	0.18	0.00	0.14	0.22	0.34	0.09	0.01	0.00	0.64	0.01	0.01	0.95	0.60	0.03
F-MIDAS_M	0.00	0.00	0.00	0.01	0.80	0.12	0.15	0.23	0.00	0.03	0.00	0.04	0.21	0.00	0.00	0.39	0.00	0.00	0.94	0.54	0.14
C-MIDAS_M	0.10	0.00	0.00	0.00	0.00	0.00	0.19	0.59	0.00	0.74	0.06	0.33	0.43	0.00	0.00	0.00	0.40	0.42	0.25	0.03	0.00
L-BVAR_M	0.74	0.36	0.26	0.00	0.02	0.22	0.32	0.16	0.57	0.09	0.50	0.69	0.00	0.03	0.01	0.08	0.24	0.06	0.26	0.17	0.20
D-BVAR_M	0.55	0.57	0.02	0.00	0.00	0.00	0.04	0.51	0.34	0.65	0.16	0.00	0.16	0.01	0.00	0.40	0.21	0.01	0.00	0.00	0.00
FADL_L	0.17	0.00	0.00	0.03	0.04	0.00	0.25	0.05	0.04	0.06	0.02	0.01	0.16	0.00	0.00	0.69	0.00	0.00	0.03	0.00	0.00
F-MIDAS_L	0.02	0.00	0.00	0.01	0.03	0.00	0.03	0.08	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.20	0.00	0.00
C-MIDAS_L	0.00	0.17	0.00	0.00	0.08	0.06	0.00	0.27	0.00	0.60	0.09	0.42	0.77	0.20	0.04	0.64	0.73	0.46	0.11	0.01	0.00
L-BVAR_L	0.00	0.00	0.01	0.00	0.02	0.05	0.09	0.10	0.00	0.09	0.04	0.00	0.03	0.00	0.00	0.20	0.03	0.00	0.26	0.05	0.03
D-BVAR_L	0.00	0.00	0.00	0.05	0.11	0.00	0.23	0.94	0.53	0.17	0.09	0.00	0.00	0.08	0.00	0.18	0.33	0.00	0.00	0.00	0.00
DSGE	0.68	0.39	0.01	0.02	0.26	0.14	0.04	0.00	0.00												

Table 4C: Annual (Y/Y) Inflation

	US			UK			Eurozone			Germany			France			Italy			Japan		
Models	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
Annual Inflation																					
AR	0.73	0.02	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.99	0.79	0.74	0.12	0.02	0.00	0.61	0.03	0.00	0.01	0.00	0.00
L-BVAR_S	0.01	0.05	0.06	0.00	0.02	0.48	0.39	0.00	0.00	0.00	0.17	0.03	0.00	0.12	0.19	0.37	0.22	0.29	0.31	0.80	0.43
D-BVAR_S	0.70	0.03	0.00	0.01	0.07	0.17	0.18	0.00	0.00	0.02	0.00	0.00	0.13	0.07	0.45	0.02	0.04	0.23	0.88	1.00	0.26
FADL_M	0.24	0.03	0.00	0.11	0.90	0.00	0.00	0.00	0.00	0.17	0.00	0.19	0.43	0.01	0.00	0.74	0.20	0.00	0.89	0.02	0.00
F-MIDAS_M	0.00	0.00	0.00	0.04	0.76	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.36	0.00	0.00	0.99	0.76	0.01
C-MIDAS_M	0.06	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.01	0.00	0.12	0.00	0.00	0.33	0.12	0.10	0.01	0.00	0.00
L-BVAR_M	0.76	0.05	0.99	0.00	0.67	0.42	0.32	0.01	0.81	0.06	0.43	0.02	0.01	0.41	0.10	0.11	0.62	0.46	0.17	0.63	0.57
D-BVAR_M	0.63	0.48	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.70	0.71	0.10	0.14	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
FADL_L	0.14	0.01	0.00	0.14	0.02	0.00	0.01	0.00	0.00	0.06	0.00	0.00	0.26	0.00	0.00	0.82	0.00	0.00	0.05	0.00	0.00
F-MIDAS_L	0.14	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.34	0.00	0.00
C-MIDAS_L	0.58	0.09	0.00	0.00	0.03	0.34	0.00	0.00	0.00	0.05	0.01	0.01	0.32	0.55	0.01	0.45	0.12	0.02	0.01	0.00	0.00
L-BVAR_L	0.00	0.00	0.00	0.00	0.65	0.68	0.10	0.68	0.00	0.11	0.40	0.00	0.03	0.16	0.00	0.25	0.23	0.00	0.31	0.11	0.00
D-BVAR_L	0.00	0.00	0.00	0.04	0.00	0.00	0.76	0.00	0.02	0.23	0.94	0.04	0.00	0.02	0.00	0.06	0.13	0.00	0.00	0.00	0.00

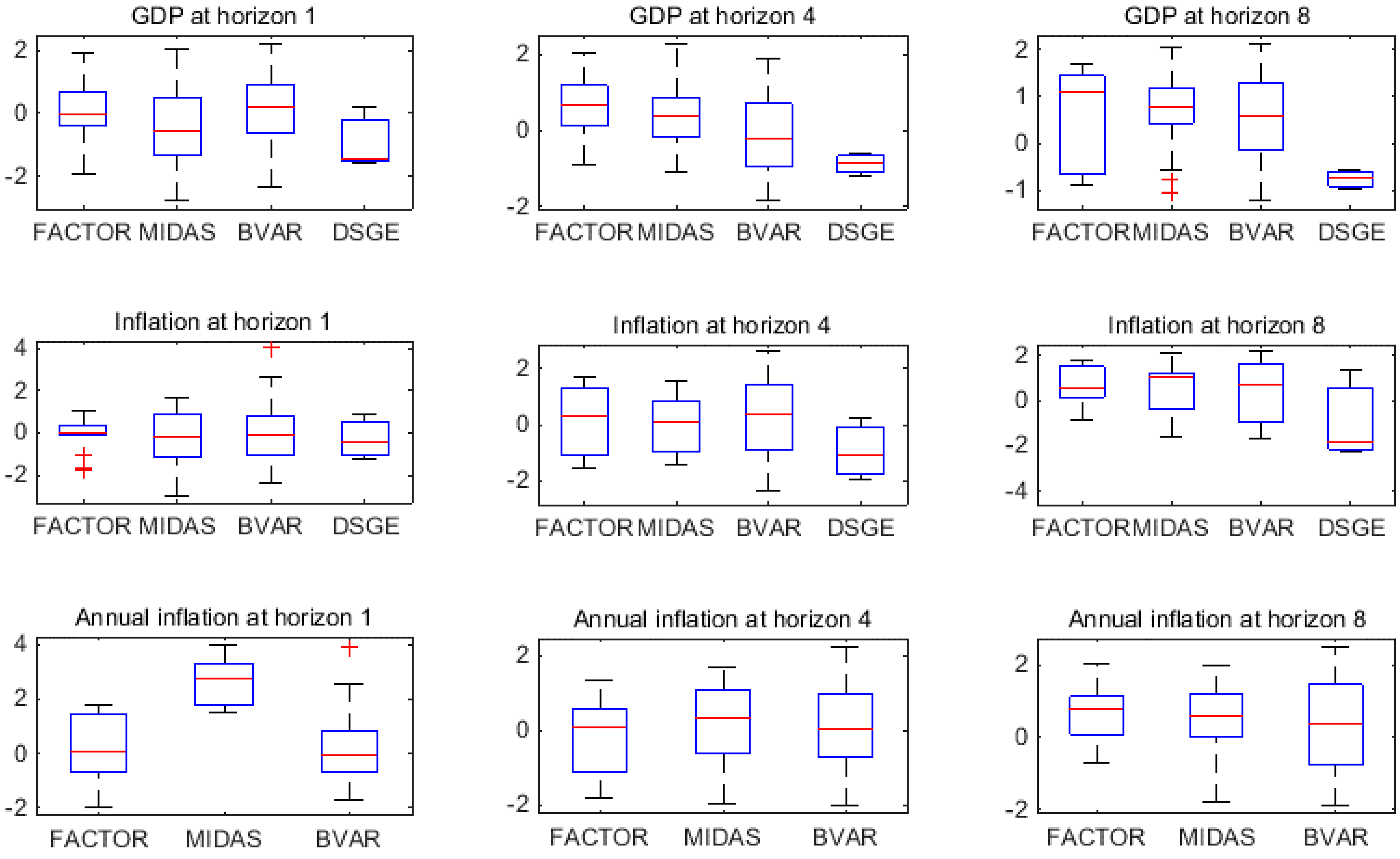
Notes: Entries are p-values for the parametric test of uniformity with no restriction imposed on serial correlation. Values in green indicate the cases that the null hypothesis of uniformity is not rejected at 10%.

Table 5: Explaining Variation in RMSFE and Logscores (relative to AR) by country, forecasting origin period, horizon, model class and dataset size.

	Rperf1 (RMSFE)			Rperf2 (MLS)		
	Out. Gr.	Q Infl	A Infl	Out. Gr.	Q Infl	A Infl
const (FADL_M, h=1	1.039***	1.015***	0.990***	1.038***	1.055***	1.032***
93-11, US+UK)	(0.013)	(0.015)	(0.017)	(0.023)	(0.021)	(0.033)
Japan	0.016***	0.068***	0.126***	0.011	-0.022**	0.171***
	(0.005)	0.010	(0.014)	(0.011)	(0.010)	(0.015)
European	-0.038***	-0.012	0.002	-0.100***	-0.066***	-0.040
(Euro, GER, IT, FR)	(0.005)	(0.008)	(0.011)	(0.008)	(0.011)	(0.023)
93Q1-97Q4	-0.081***	0.130***	0.239***	0.014	0.118***	0.091
	(0.008)	(0.015)	(0.022)	(0.009)	(0.014)	(0.051)
98Q1-02Q4	-0.046***	-0.004	0.012	0.025***	-0.032***	-0.019
	(0.005)	(0.009)	(0.013)	(0.009)	(0.012)	(0.019)
03Q1-07Q4	0.002	0.011*	-0.015*	-0.004	-0.014	-0.092***
	(0.005)	(0.006)	(0.008)	(0.013)	(0.012)	(0.021)
08Q1-11Q3	0.043***	0.012	-0.014	0.072***	-0.008	-0.085***
	(0.007)	(0.009)	(0.011)	(0.012)	(0.013)	(0.018)
h=2,…,4	-0.040***	-0.012	0.014	-0.002	-0.028	-0.043
	(0.013)	(0.015)	(0.017)	(0.022)	(0.021)	(0.027)
h=5,..,8	-0.028**	-0.035**	-0.039**	-0.065***	-0.057**	-0.104***
	(0.013)	(0.017)	(0.020)	(0.025)	(0.023)	(0.033)
MIDAS	0.029	-0.002	-0.334***	0.011	0.012	-0.463***
	(0.018)	(0.019)	(0.021)	(0.028)	(0.027)	(0.043)
BVAR	-0.013	-0.011	-0.031	0.004	-0.098***	-0.131***
	(0.016)	(0.019)	(0.021)	(0.028)	(0.027)	(0.033)
(h=2,…,4)*MIDAS	-0.021	-0.033	0.203***	-0.035	-0.051*	0.276***
	(0.019)	(0.021)	(0.024)	(0.029)	(0.030)	(0.040)
(h=5,…,8)*MIDAS	-0.060***	-0.006	0.343***	-0.048	-0.024	0.350***
	(0.020)	(0.023)	(0.027)	(0.033)	(0.032)	(0.060)
(h=2,…,4)*BVAR	0.058***	-0.013	0.012	-0.025	-0.007	0.072**
	(0.017)	(0.021)	(0.023)	(0.028)	(0.029)	(0.035)
(h=5,…,8)*BVAR	0.024	0.006	0.041	-0.019	0.028	0.144***
	(0.017)	(0.022)	(0.026)	(0.030)	(0.030)	(0.039)
Small	-0.031***	0.027**	0.038***	-0.005	0.002	0.010
	(0.006)	(0.011)	(0.015)	(0.011)	(0.012)	(0.019)
Large	-0.010	-0.042***	-0.055***	-0.005	-0.015	-0.051*
	(0.008)	(0.015)	(0.019)	(0.017)	(0.019)	(0.030)
Large* BVAR	-0.029**	-0.070***	-0.063**	-0.164***	-0.201***	-0.156***
	(0.012)	(0.020)	(0.025)	(0.021)	(0.024)	(0.036)
Large*MIDAS	0.001	0.021	0.024	-0.013	-0.030	-0.035
	(0.011)	(0.020)	(0.025)	(0.024)	(0.026)	(0.055)
R^2	0.151	0.126	0.211	0.156	0.190	0.092
Nobs	2976	2976	2976	2976	2976	2976
Mean of dep. var.	0.978	0.986	0.969	0.929	0.888	0.810

Note: Values larger than 1 imply that model improves over the AR. All explanatory variables are dummy variables. Regressions estimated by OLS. Values in brackets are White standard errors. *,**,*** denote rejection of the null of no statistical accuracy, respectively, at 10%, 5% and 1%.

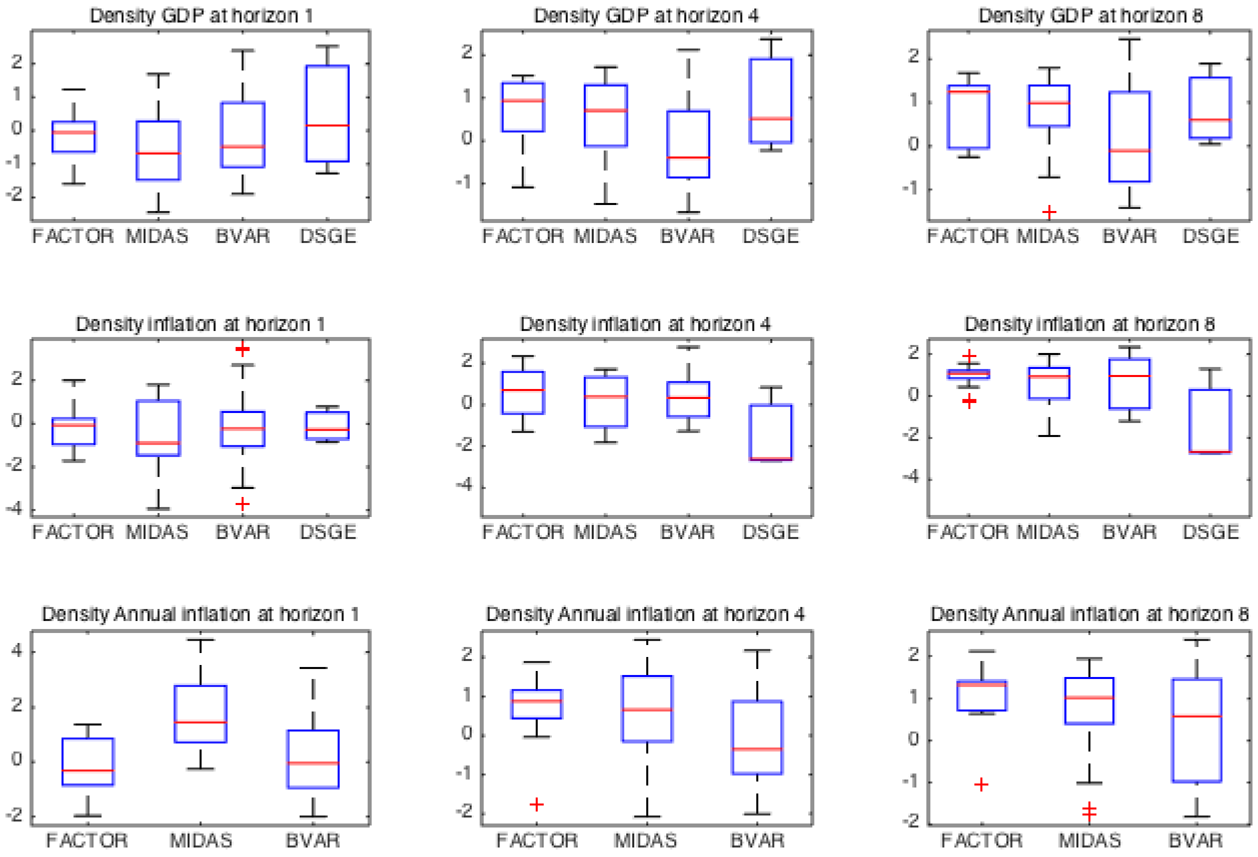
Figure 1: Characteristics of the Distribution of the equal accuracy t-statistic for differences in MSFE, computed over full out-of-sample, with the AR model under the null and a forecasting model from the model class indicated under the alternative (aggregate over specifications and countries).



Note: “On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.” (Matlab description). VAR specifications included L-BVAR_S, L-BVAR_M, L-BVAR_L, D-BVAR_S, D-BVAR_M, D-

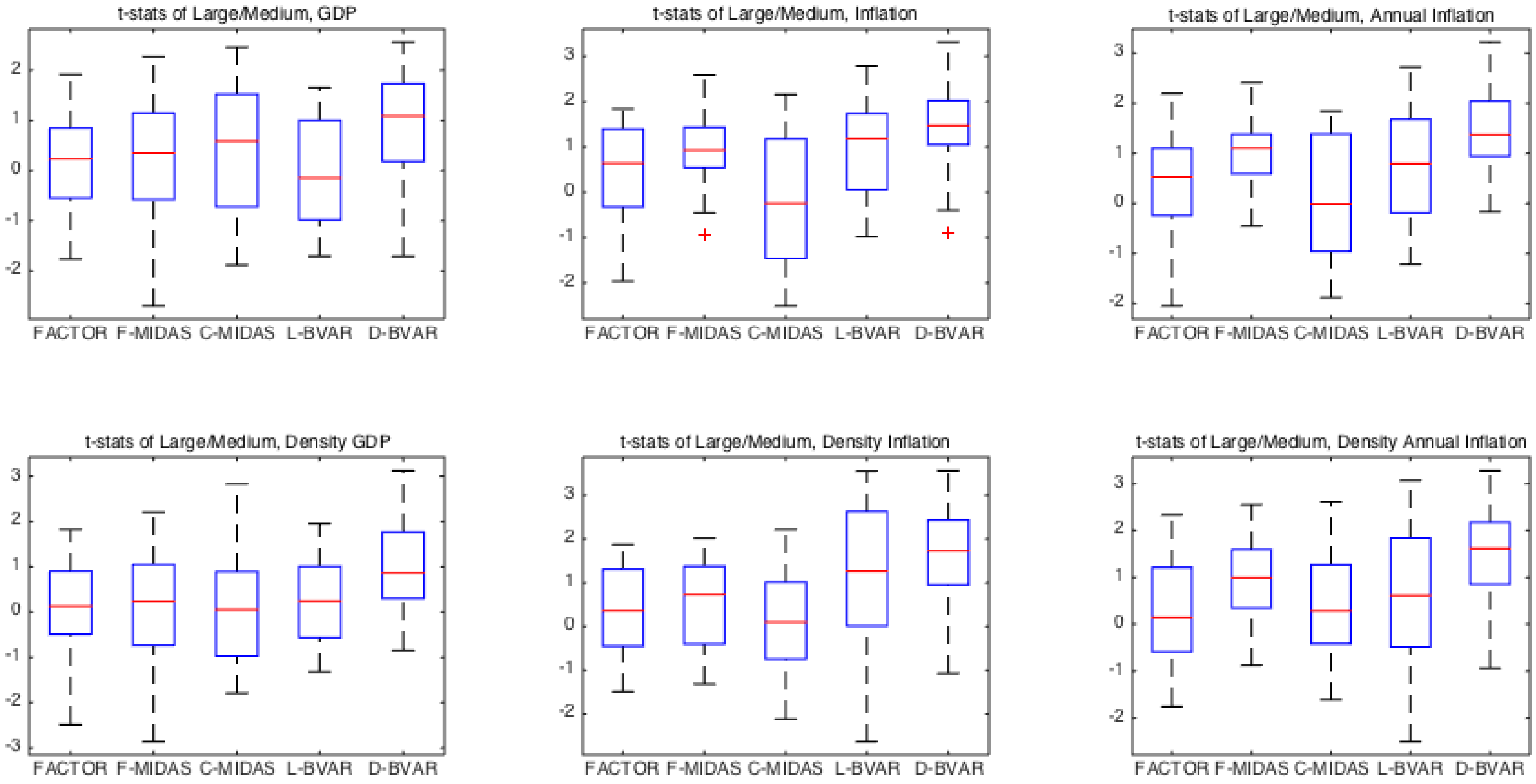
BVAR_L. MIDAS specifications include F-MIDAS_M, C-MIDAS_M, F-MIDAS_L, C-MIDAS_L. Factor Specifications include FADL_M and FADL_L. Only one DSGE specification for 3 countries. All other models, the results are for all 7 countries.

Figure 2: Characteristics of the Distribution of the equal accuracy t-statistic for differences in logscore, computed over full out-of-sample, with the AR model under the null and a forecasting model from the model class indicated under the alternative (aggregate over specifications and countries).



Notes: See notes of Figure 1.

Figure 3: Characteristics of the Distribution of the equal accuracy t-statistic for a model with a medium dataset against a large dataset, computed over full out-of-sample, for each indicated forecasting model (aggregate over forecasting horizons (1 to 8) and countries (7)).



Note: Each box plot is based on h=8*countries=7= 56 values. See notes of Figure 1. In the first panel , the accuracy test is based on MSFEs, and in the second panel, the statistics are based on the logscore.