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

August, 2018

Abstract

By employing datasets for seven developed economies and considering four classes of multivariate forecasting models, we extend and enhance the empirical evidence in the macroeconomic forecasting literature. The evaluation considers forecasting horizons from one-quarter up to two-years ahead. We find that the structural model, a medium-sized DSGE model, provides accurate US and UK long-horizon inflation forecasts. To strike a balance between being comprehensive and producing clear messages, we employ meta-analysis regressions to 2,976 relative accuracy comparisons that vary with forecasting horizon, country, model class and specification, number of predictors, and evaluation period. For point and density forecasting of GDP growth and inflation, we find that models with a large number of predictors do not perform better than models with 13-14 hand-picked predictors. Factor-augmented models and equal-weighted combinations of single-predictor mixed-data sampling regressions are a better choice for dealing with a large number of predictors than Bayesian VARs.

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

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 research assistant Katerina Petrova, who was 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. As a result, considerable efforts have been made in academic work to lay the foundations and build tools for efficient forecasting.

The macroeconomic forecasting literature can be divided into two broad categories. 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 stream of research 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 makes it harder to provide economic storytelling to support forecasts. This characteristic is classified as a relevant disadvantage by many 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 necessarily correlate closely with superior forecasting performance.[2] The second trend of considering large datasets has been spurred by their use in many economic analyses, given their availability in central banks and other policy making institutions.[3]

[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.

[3] Stock and Watson (2002) is an influential paper supporting the use of large datasets for forecasting macroeconomic variables. Other more recent contributions, all pointing towards the importance of using medium-large dataset for

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 model classes for forecasting output growth and inflation, giving special attention to model classes able to deal with a large number of predictors. The aim 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 aim to compare and contrast reduced-form models and structural models, which have traditionally been considered inferior for forecasting purposes. This latter aspect of our analysis is less commonly found in forecasting evaluations.[4]

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 instead data from seven economies: US, UK, Euro Area, Germany, France, Italy and Japan. For these seven economies, we compute forecasts for output growth and inflation with three classes of state-or-art reduced-form forecasting models: Factor-Augmented Distributed Lag (FADL) Models, Mixed Data Sampling (MIDAS) Models, Bayesian Vector Autoregressive (BVAR) Models.[6] These model classes are useful to explore the

macroeconomic forecasting, include Bańbura, Giannone and Reichlin (2010), Carriero, Clark and Marcellino (2015), Koop (2013) and Giannone, Lenza and Primiceri (2015).

[4]Density forecasts of DSGE models are evaluated by Del Negro and Schorftheide (2013) and Diebold, Schorftheide and Shin (2017), 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 et al. (2015) evaluate models for 19 countries, but they use only a relatively small set of predictors.

[6]Time-varying vector autoregressive models, expoited as forecasting models by D'Agostino, Gambetti and Giannone (2013), and vector autoregressive models with stochastic volatility, with forecasting performance evaluated by Clark (2011), are classes of models that are excluded from this forecasting comparison. The main reason is that both classes are not easily adaptable to large datasets. The proposed approach by Koop and Korobilis (2013) for large datasets considers a VAR with 25 variables as "large". In this paper, we use datasets up to 155 variables. We also use data from contries with shorter time series where structural changes are harder to identify. In this paper, we consider just one class of mixed frequency models. Mixed frequency specifications are popular for nowcasting as surveyed by Banbura, Giannone, Modugno and Reichlin (2013), including recent contribution by Schorfheide and Song (2015). Because we aim to evaluate forecasting performance from nowcasting up to long horizons (two years), we select just one class of

predictive information of a large number of indicators. As a consequence, we build a dataset with a large number of monthly indicators for each country and assess the importance of employing large (one-hundred predictors) datasets in comparison with medium-sized (a dozen predictors) and small datasets in macroeconomic forecasting. We also consider one class of structural models: a medium-sized Dynamic Stochastic General Equilibrium Model (DSGE). We compare the DSGE performance with reduced-form models for forecasting output growth and inflation in the US, the UK and the Euro area.

We have some knowledge of the relative point 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 et al. (2013)), and of Bayesian VARs to Dynamic Factor models (Bańbura et al. (2010)). In this paper, we advance further by comparing the out-of-sample forecasting accuracy for point and density forecasts of output growth and inflation for the following class of models: BVAR, FADL, MIDAS and DSGE models.

The design of our forecasting comparison with the elements described above imply that we evaluate the forecasting performance of 13 reduced-from model specifications to predict two quarterly macroeconomic time series over horizons from one-quarter to eight-quarters ahead. And we do this comparison for seven different countries and consider four different subperiods of 5 years over a 20 year out-of-sample period. In order to get clear messages from our empirical exercise, we develop evaluation methods that pool forecasting performance across countries, model class, forecasting origin period and dataset size.

Our meta-analysis method employs a regression of the relative performance of each multivariate reduced-form model on a set of characteristics. The relative performance is measured using the root mean squared forecast error for point forecasts and logscores for density forecasts. The performance is measured with respect to the autoregressive model for the same variable and horizon. The method allows us to assess the statistical significance of forecasting horizon, geographical source (country), model class, evaluation period and number of predictors (dataset size) in explaining forecasting performance.

A second evaluation method relies on t-statistics for a Diebold and Mariano (1995) equal forecast accuracy test for the 20-year evaluation period. We investigate the empirical distribution of t-statistics with an autoregressive model under the null. We use this approach to complement the results of the meta-analysis when comparing the point and density forecasting performance of specifications that use a large set of predictors in comparison with the ones that use a smaller set. We also use empirical distributions of equal accuracy t-statistics against an AR benchmark to evaluate

mixed frequency models that has relatively good nowcasting performance (Andreou, Ghysels and Kourtellos (2013) and Kuzin et al. (2013)).

how the structural models forecasting accuracy compares with reduced-form models.

We find no support for the use of large datasets (one-hundred predictors) instead of medium-sized (a dozen predictors) ones. However, we provide evidence that the factor model and an equal-weighted combination of single regressor MIDAS models are the best specifications to deal with large datasets since they perform on average better than Bayesian VARs. We find that DSGE models have relative good performance for forecasting US and UK inflation at forecasting horizons longer than one year.

The empirical results provide only limited support to the use of mixed frequency models, which exploit current quarter information on monthly series, to improve nowcasts of output growth. The reason is that there is large cross-country variation on nowcasting performance of mixed frequency models. The results also suggest changes in the relative forecasting performance of forecasting models. The relative performance of reduced-form multivariate models is at its peak in the 1993-1997 period for inflation and in the 2008-2011 period for output growth.

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 key elements of the design of our forecasting exercises, including statistical tests employed. In section 5, we explore the key determinants of point and density macroeconomic forecasting performance of multivariate statistical models to AR models using meta-analysis regressions and the empirical distribution of equal-accuracy t-statistics. An evaluation of the point and density forecasting accuracy of structural models in comparison to reduced-from models is discussed in section 6. Section 7 concludes.

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. Density forecasting evaluation provides us with insights on the accuracy of forecasting models for the whole predictive distribution. 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.

In the remainder 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 forecasts from univariate autoregressive (AR(p)) models. The autoregressive order is selected 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 $\Delta \hat{q}^{(i)}_{T+1}, ..., \Delta \hat{q}^{(i)}_{T+h_{\max}}$ 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,k) equation for each horizon h:

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

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 B2 and B3 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 $k = 1$ 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 k using the modified SIC, and normally $k = 1$ 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 (2016), 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).[7] 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.

2.3 MIDAS Models

The economic predictors in our dataset, summarized in Table 2, 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:

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

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 k, and km 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 et al. (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 utilization of monthly data is the main advantage of the MIDAS approach for macroeconomic forecasting (Clements and Galvão, 2008; Kuzin et al., 2013; Andreou et al., 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

[7] 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.

sum up to one. We use the beta function to obtain the weights, that is,

$$w(\theta; i) = \frac{f(\theta; i)}{\sum_{j=1}^{K} f(\theta; j)}$$
$$f(\theta; i) = \frac{(j)^{\theta_1 - 1}(1-j)^{\theta_2 - 1}\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}; \;\; j = i/km.$$

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 et al. (2016) 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 employ 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 B2 and B3 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 and to two in the case of large datasets following Andreou et al. (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 et al. (2015), 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:

$$y_t = A_0 + A_1 y_{t-1} + ... + A_p y_{t-p} + \varepsilon_t \tag{2}$$

$$\varepsilon_t \sim N(0, \Sigma)$$

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

We elicit a conjugate Normal-Inverse Wishart prior:

$$\alpha|\Sigma \sim N(\alpha_0, \Sigma \otimes \Omega_0)$$

$$\Sigma \sim IW(S_0, v_0),$$

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

$$\alpha|\Sigma, data \sim N(\overline{\alpha}, \Sigma \otimes \overline{\Omega})$$

$$\Sigma|data \sim IW(\bar{S}, \bar{v}).$$

Carriero et al. (2015) describe the close form solutions for the posterior means and variances, and the prior mean and variances under the assumption that they follow a Minnesota-style prior as in Bańbura et al. (2010). We consider prior means for the first-order autoregressive coefficients equal to one if the endogenous variables, y_t, are in log-levels as described above. We also consider a specification in differences, using Δy_t, with the prior mean equal to zero.

We also impose -in the case of VAR in levels- the sum of coefficients prior, which expresses the belief that the average of the past values of a given variable provides a good forecast for that variable. The fact that, in the limit, the sum-of-coefficients prior is not consistent with cointegration motivates the use of an additional prior, known as the 'dummy initial observation' prior. 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). These last two priors together tend to improve forecasts when dealing with data in levels. Hyperparameters governing priors are set as the baseline case in Carriero et al. (2015). The overall prior tightness λ_1 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 et al. (2015). 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 the procedure is split 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. 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 $p = 4$. When the target forecasting variable is the quarterly growth rate, we transform accordingly the forecasts for the model in levels.

2.5 DSGE Models

The literature provides evidence of accuracy of the medium-sized Smets and Wouters (2007) model (Christoffel, Coenen and Warne, 2010; Edge and Gurkaynak, 2011; Del Negro and Schorftheide, 2013; Woulters, 2015). 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 equally-spaced draws from the kept posterior parameters draws generated by the MCMC procedure 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 B1 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.

For each country, we build a medium and a large dataset of economic indicators sampled monthly. The datasets are summarized in Table 2 and described in detail in Tables B2 and B3 of the online appendix. When quarterly data are required, we use the average over quarter for factor models, so F-MIDAS nest FADL models[8], 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 et al. (2013) datasets. The medium dataset includes 11-14 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 et al. (2015). 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 as recorded in Table 2. It varies between 57 (Japan, France) and 155 (US). The large dataset includes also all variables in the medium dataset. 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. We provide the description of all variables including their datastream code in the Table B3 in the online appendix.[9]

Because of the lack of availability of real-time dataset for the monthly indicators for all our seven countries, we use only data from the currently available vintage as it is generally the case when evaluating forecasts with models for large datasets (as, for example, Smets and Wouters (2007) and Kuzin et al. (2013)).

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 B4 in the online appendix describes the variables employed in the DSGE estimation, including their required

[8] This implies that F-MIDAS specification nests the FADL if the MIDAS weighting function is flat, that is, $\theta_1 = \theta_2 = 1$.

[9] Some variables were seasonally adjusted by the X12 filter before estimation, and they have SA indicated in Table B3.

transformation.

The last observation employed in our forecasting exercise is 2013M9. For US, Japan and UK, we use data from 1975M1 (with exception of UK CPI inflation which is only available from 1980M1), but for other countries, data is only available later as described in Table 2. Data for DSGE estimation is from 1984Q1 for the US, UK and the 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.

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. The accuracy of point forecasts is measured by Root Mean Squared Forecast Errors (RMSFE), and the log predictive score measures the accuracy of density forecasts. The advantage of using log scores to compare density forecasts is that the maximization of the logscore is equivalent to minimize the Kullback-Leibler distance between the model and the true density. To compute log scores, we first fit a Gaussian kernel density to the 5000 predictive density draws over a grid between -15 and 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 for equal accuracy. The variance is computed with the Newey-West estimator with maximum order increasing with the horizon.

Table 1 provides a short description of all forecasting models we employ in this evaluation. Similarly to Bańbura et al. (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 Explaining forecasting performance of statistical models

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 one structural model (DSGE). In this section we explore the relative forecasting performance of the 12 multivariate reduced-form models, listed as models 2 to 13 in Table 1. Forecasting comparisons that include the DSGE model are discussed in section 6. We measure the impact of model class, forecasting horizon, dataset size and data source (country) on point and density forecasting performance.

5.1 A Meta Analysis

Our aim is to investigate how the relative (to the AR model) forecasting performance of each statistical model class (MIDAS, FADL and BVAR) varies with the number of predictors (medium vs large dataset), the forecasting horizon (nowcasting, short-horizon ($h = 2, .., 4$) and medium-horizon ($h = 5, ..., 8$), the 5-year subperiod evaluated, and the geographical source of the dataset.

The dependent variable in our meta analysis regression is a measure of the relative forecasting performance of a specific forecasting model to the autoregressive model when predicting one of the target variables (output growth and 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)[10] computed for a specific target variable varying across country, forecasting model, period and horizon. The measures for point and density forecasting performance are:

$$rMSFE_{m,p,c,h} = \frac{RMSFE_{AR,p,c,h}}{RMSFE_{m,p,c,h}};$$

$$rMLS_{m,p,c,h} = 1 + [(-MLS_{ar,p,c,h}) - (-MLS_{m,p,c,h})].$$

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 $h = 1, ..., 8$.

As consequence, the total number of relative performance observations (considering that the forecasting period availability varies across countries as noted in Table 2) is 2,976. By exploiting a large set of forecasting comparisons, we aim to find sources of performance improvements in macroeconomic forecasting that are not constrained by model class, forecast horizon, country and evaluation period.

[10] We use the median instead of the mean logscore to minimize the impact of outliers in our analysis. Outlier values are more frequent with logscores than with squared forecast errors.

The first characteristic we explore is the country where data is sourced. We use two dummy variables to split the country set in Table 2 into three: $D^{EU} = 1$ for Euro Area countries (c =EU, FR, IT, GER) (and $D^{EU} = 0$ otherwise), and $D^{JP} = 1$ if c =JP. The benchmark countries are then US and UK.

The second characteristic is the forecasting horizon. We split the set of forecast horizons into three groups by defining $D^{sh} = 1$ if $h = 2, ..., 4$ and $D^{mh} = 1$ if $h = 5, ..., 8$. Accordingly, differences in performance over short and medium horizons are assessed against the nowcasting ($h = 1$) benchmark.

We are also interested in finding differences between the three model classes. We set 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 the number of predictors 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 dataset size is the medium one.

Finally, the impact of the evaluation period is assessed by creating one dummy variable for each one of the four five-year out-of-sample subperiods. As a consequence, performance improvements are relative to the full out-of-sample (p =93Q1-11Q3).

We also consider interactions between the dummy variables described above. We consider interactions between horizon and model class dummies, between $D^{l\arg e}$ and model class dummies, and between $D^{l\arg e}$ and evaluation period dummies.

The meta-analysis regression is then:

$$
\begin{aligned}
rLoss_{m,p,c,h} = {} & \beta_0 + \beta_1 D^{JP} + \beta_2 D^{EU} \\
& + \beta_3 D^{9397} + \beta_4 D^{9802} + \beta_5 D^{0307} + \beta_6 D^{0811} \\
& + \beta_7 D^{sh} + \beta_8 D^{lh} + \beta_9 D^{MIDAS} + \beta_{10} D^{BVAR} \\
& + \beta_{11} D^{MIDAS} * D^{sh} + \beta_{12} D^{MIDAS} * D^{lh} + \beta_{12} D^{BVAR} * D^{sh} + \beta_{14} D^{BVAR} * D^{lh} \\
& + \beta_{15} D^{small} + \beta_{16} D^{l\arg e} + \beta_{17} D^{l\arg e} * D^{BVAR} + \beta_{18} D^{l\arg e} * D^{MIDAS} \\
& + \beta_{19} D^{l\arg e} * D^{9397} + \beta_{20} D^{l\arg e} * D^{9802} + \beta_{21} D^{l\arg e} * D^{0307} + \beta_{22} D^{l\arg e} * D^{0811} + \varepsilon_{m,p,c,h}.
\end{aligned} \tag{3}
$$

for $m = 2, ..., 13; p =$ 93-97,98-02,03-07,08-11,93-11; $h = 1, ..., 8; c = US, UK, JP, FR, IT, GER, EU$

$rLoss_{m,p,c,h}$ is either $rMSFE_{m,p,c,h}$ or $rMLS_{m,p,c,h}$.

Note that β_0 measures the relative (to the AR model) performance of the the FADL medium model ($m = 2$) for $h = 1$ over the full sample period (p =93-11) with US and UK data ($c = 1, 2$). As consequence, all other coefficient estimates are measures of gains/losses against this benchmark.

5.2 Meta-Analysis Results

Table 3 presents estimates of the regression in (3) with standard errors clustered by country, implying that we consider country-specific effects. The table columns describe results for each performance measure (*rMSFE* and *rMLS)* and target variable (output growth, inflation). Cases where the null hypothesis that the coefficient is equal to zero is rejected are indicated with stars for 10%, 5% and 1% significance levels. Values in bold show estimates are statistically significant at 10% when using heteroscedasticity-robust standard errors instead of the country-clustered standard errors displayed in Table 3.

The characteristics considered in regression (3) explain between 13% and 20% of the forecasting performance depending on the target and the type of performance measure. As a consequence, idiosyncratic variation has an important role in explaining forecasting performance across this large number of forecasting exercises. The following analysis will consider characteristics with statistically significant role in explaining forecasting performance, as indicated in Table 3.

The estimates of the regressions' intercepts are all larger than 1, implying that on average the FADL_M improves over the AR when nowcasting US and UK variables. Gains are larger for output growth and imply a 4% improvement in RMSFE. Estimates for β_1 and β_2 suggest that benefits of employing multivariate models instead of AR models for predicting output growth are larger with Japanese data but smaller with European data.

The estimated coefficients on the evaluation period dummies point to changes in statistical performance over time, but the estimates are statistically significant with country-clustered standard errors only when evaluating output growth point forecasts. During the turbulent 08Q1-11Q3 period, we find that multivariate models perform relatively better for output growth, but they do relatively worse in the 98Q1-11Q3 period.

The estimated coefficients on the forecasting horizon dummies are all negative, implying that the relative performance of multivariate models to the AR model deteriorates with the horizon. This deterioration is statistically significant for point forecasting output growth and inflation when horizon is iterated with the MIDAS model dummy variable. This declining MIDAS forecasting performance with horizon is partially compensated by the fact that MIDAS models improve RMSFEs over the benchmark in 3% on average when nowcasting output growth, albeit the estimate of β_9 is not statistically significant. For predicting output growth, BVAR models do relatively better at medium horizons ($h = 5, ..., 8$) and are significantly better at $h = 2, 3, 4$. These results suggest that although MIDAS models may deliver accurate nowcasts of output growth for some countries, this class of models performance deteriorates rapidly with the forecast horizon and a BVAR specification may be a more accurate choice in some cases.

The estimated coefficients on the dataset-size dummies indicate that BVARs with only three

variables, including both targets, are significantly worse than models with a moderate number of indicators in predicting output growth. For predicting inflation, either a small or a medium set of indicators perform significantly better than large datasets.

The interactions between dataset size and model class clearly indicate that large BVAR models deteriorate forecasting performance. These results suggest that if the aim is to exploit information in a large number of predictors (more than 55 indicators) for forecasting output growth and inflation, then the use of models with factors (FADL and F-MIDAS) or forecasting combinations (C-MIDAS) are more adequate than BVAR models. However, there is no evidence that the use of a large number of predictors instead of a dozen picked variables (medium dataset) improves macroeconomic forecasting. By evaluating the estimates for the iterations between $D^{l\,\mathrm{arg}\,e}$ and the sample period, we find that a large set of predictors worsens output growth point forecasting performance in the earlier periods when sample sizes employed in the estimation are shorter (recall that we increase sample size when estimating models at each forecasting origin).

In summary, we find some time variation in the relative forecasting performance of multivariate statistical models to AR models for forecasting output growth across countries: multivariate models are in particularly useful during the last four year period (2008-2011). We find no evidence that models with a larger number of predictors improve over the performance of models with smaller set of predictors. If using a large dataset, FADL and MIDAS models are more adequate than BVAR models. We find very limited evidence that MIDAS models improve nowcasts.

5.3 Additional Meta-Analysis Comparisons

For MIDAS and BVAR model classes, we consider two main specification types. For MIDAS models, we compute forecasts by using a factor-augmented version (F-MIDAS) and an equal-weight forecasting combination strategy (C-MIDAS). For BVAR models, we use a specification in levels (L-BVAR) and another in growth rates (D-BVAR). In this subsection, we use relative performance regressions to test if there are any statistical differences in performance between these specification types that hold across countries, horizon, evaluation period and number of predictors.

In Table 4A, we present results for the four measures of performance in Table 3 (output growth and inflation; *rMSFE* and *rMLS*). These are single regressions estimated with performance measures computed only for MIDAS models ($rLoss_{m,p,c,h}$ for $m = 4, ..., 7$ with p, h and c variation as in (3)). We define the dummy variable D^{CMIDAS} as equal to 1 if $m = 6, 7$. As a consequence, if the estimated coefficient of D^{CMIDAS} is significantly positive, we can conclude that the equal-weighted forecasting combination of single regressor MIDAS models is a better way to exploit the information on a set of predictors than using monthly factors. The coefficients are indeed positive and statistically significant with country-clustered standard errors in all columns of Table 4A, so we conclude in favour of the

C-MIDAS specifications.

In Table 4B, we compute single regressions with the same performance measures, but for BVAR models only ($rLoss_{m,p,c,h}$ for $m = 8, ..., 13$ with p, h and c variation as in (3)). We define the dummy variable D^{DBVAR} as equal to 1 if $m = 9, 11, 13$ and zero otherwise. The empirical results can inform us on whether the BVAR-in-differences improves over the BVAR-in-levels. Recall that the main advantage of using the BVAR-in-levels (L-BVAR) is that the possibility of cointegration is allowed for. The results in Table 4B suggest that this BVAR specification choice only matters for point forecasting output growth: L-BVARs perform significantly better than D-BVARs.

5.4 Evaluating the impact of the dataset size with equal accuracy tests

Our previous results suggest that the use of forecasting models with a large set of predictors may have a negative effect on forecasting performance for both output growth and inflation, in particularly if using BVAR models with short samples. In this subsection, we evaluate this research question using the empirical variation of "medium vs large" equal accuracy tests for point and density forecasts as described in section 4.

Figure 1 presents empirical t-statistics distributions for the following models: FADL, F-MIDAS, C-MIDAS, L-BVAR and D-BVAR. The Diebold and Mariano (1995) t-statistics are computed with the specification with a medium dataset under the null and the model with the large dataset under the alternative using the full out-of-sample period (p =93-11). The box plots are computed for t-statistics obtained for different horizons ($h = 1, ..., 8$) and countries. Negative values imply that the model with a large number of predictors is more accurate than the same model 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.

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. However, based on the median t-statistics, we can say that D-BVARs are worse in handling large datasets than L-BVARs, providing an additional nuance to our results in section 5 discouraging the use of BVARs with large datasets. These results also support the use of the C-MIDAS specification instead of the F-MIDAS in particularly when dealing with large datasets for forecasting inflation.

In summary, there is no strong evidence that a large number of predictors improve forecasts over a moderate ammount, but we can provide evidence to support the use of C-MIDAS and FADL specifications to deal with large datasets instead of BVAR models.

6 Comparing structural vs reduced-form forecasting models

In the previous section, we investigate common features that explain relative forecasting performance of reduced-form statistical models across countries, forecasting horizons, forecasting periods and model specification. In this section, we use equal accuracy tests computed as described in Section 4 to compare the performance of reduced-form statistical models (FADL, BVAR, MIDAS) with the DSGE model.

Details of the DSGE model employed including our estimation strategy were discussed in section 2.4. We describe the dataset employed in the estimation of DSGE models in section 3. One should note that the medium-sized DSGE forecasts are considered only for $c = US, UK, EU$ and they are estimated with output growth per person and GDP deflator inflation. To measure the relative performance of DSGE models to the AR benchmark, we recompute AR forecasts using the same measurements of output growth and inflation employed by the DSGE model.

Figures 2 and 3 present box plots of the Diebold and Mariano (1995) t-statistics. The t-statistics are computed for the full out-of-sample of period for each country as listed in Table 2. 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 empirical distributions vary with the country and are computed for a specific model class (FADL, MIDAS, BVAR, DSGE). The box plots are presented separately for three horizons ($h = 1, 4$ and 8). Figure 2 presents results for output growth and inflation using the quadratic loss function (MSFE) to compute the t-statistics. The plots in Figure 3 instead are based on the differences in logscore.

The results in Figures 2 and 3 help us to indicate which model class, including statistical model classes (FADL, MIDAS, BVAR) and the structural model class (DSGE), performs best for each target variable and for a set of forecasting horizons. The median t-statistic in Figures 2 and 3 can be employed to evaluate how each class of model performs on average across specifications and countries for each horizon and target variable.

MIDAS models do better at $h = 1$ for output growth, but the distribution of t-statistics has a large spread, suggesting that mixed frequency models improve output growth nowcasts for the median country but does not perform well for some countries. For $h = 4$, it is clear that BVARs perform better for forecasting output growth. When forecasting inflation, the clear evidence we have is that DSGE models do better when predicting inflation at $h = 4, 8$ for both point and density forecasts. The results in Figures 2 and 3 suggest that DSGE models are able to significantly improve AR forecasts of quarterly inflation at $h = 4, 8$.

These results are supported by detailed Tables by country and forecasting horizon in the online appendix. Table A1 shows the relative performance of the DSGE model against the AR and the FADL_M using RMSFEs and Table A2 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). 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).

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

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 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 one and two-year ahead forecasts of inflation not only for the US but also for the UK.

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. We propose meta-analysis regressions to be able to draw a small set of clear messages from 2,976 relative accuracy comparisons.

We extend results based only on Bayesian VARs (Koop, 2013) by showing that the use of a large set of predictors instead of a moderate set do not improve forecasts. Our contribution is to employ five different specifications from three model classes to address whether it is worth to use large datasets instead of using 10-15 chosen predictors for both point and density forecasting, and we find that indeed a medium dataset typically suffices. When dealing with a large number of predictors (more than 50) to estimate a forecasting model over a short time period, we find that factor augmented distributed lag models and equal-weighted combinations of single-predictor mixed-data sampling regressions perform better than BVARs in predicting key macroeconomic variables when considering point and density forecasting.

References

Aastveit, K. A., Foroni, C. and Ravazzolo, F. (2016). Density forecasts with MIDAS models, *Journal of Applied Econometrics* **32**: 783–801.

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.

Banbura, M., Giannone, D., Modugno, M. and Reichlin, L. (2013). Now-casting and the real-time data flow, *Handbook of Economic Forecasting, volume 2A*, Elsevier, chapter 4, pp. 195–237.

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

Carriero, A., Clark, T. E. and Marcellino, M. (2015). Bayesian VARs: Specifications choices and forecast accuracy, *Journal of Applied Econometrics* **30**: 46–73.

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* .

Christoffel, K., Coenen, G. and Warne, A. (2010). Forecasting with DSGE models, *ECB Working Paper Series n. 1185* .

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

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

D'Agostino, A., Gambetti, L. and Giannone, D. (2013). Macroeconomic forecasting and structural change, *Journal of Applied Econometrics* **28**: 82–101.

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. (2017). Real-time forecast evaluation of DSGE models with stochastic volatility, *Journal of Econometrics* **(in press)**.

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.

Giannone, D., Lenza, M. and Primiceri, G. E. (2015). Prior selection for vector autoregressions, *Review of Economic and Statistics* **97**: 412–435.

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.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter vars, *Journal of Econometrics* **177**: 185–198.

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.

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency VAR, *Journal of Business and Economic Statistics* **33**: 366–380.

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.

Woulters, M. H. (2015). Evaluating point and density forecasts of DSGE models, *Journal of Applied Econometrics* **30**: 74–96.