International Journal of Forecasting 39 (2023) 1820–1838

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Macroeconomic forecasting in the euro area using predictive combinations of DSGE models☆

Jan Čapek [a], Jesús Crespo Cuaresma [b,a,c,d,*], Niko Hauzenberger [e,a], Vlastimil Reichel [a]

[a] Masaryk University, Czech Republic
[b] Vienna University of Economics and Business, Austria
[c] IIASA, Austria
[d] WIFO, Austria
[e] University of Salzburg, Austria

ARTICLE INFO

Keywords:
Forecasting
Model averaging
Prediction pooling
DSGE models
Macroeconomic variables

ABSTRACT

We provide a comprehensive assessment of the predictive power of combinations of dynamic stochastic general equilibrium (DSGE) models for GDP growth, inflation, and the interest rate in the euro area. We employ a battery of static and dynamic pooling weights based on Bayesian model averaging principles, prediction pools, and dynamic factor representations, and entertain six different DSGE specifications and five prediction weighting schemes. Our results indicate that exploiting mixtures of DSGE models produces competitive forecasts compared to individual specifications for both point and density forecasts over the last three decades. Although these combinations do not tend to systematically achieve superior forecast performance, we find improvements for particular periods of time and variables when using prediction pooling, dynamic model averaging, and combinations of forecasts based on Bayesian predictive synthesis.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Dynamic stochastic general equilibrium (DSGE) models have become the workhorse of modern macroeconomic research, due to their internal consistency and their ability to assess the effects of policy shocks in a rigorous manner.

In spite of their importance in modern economic analysis, the existing results concerning their out-of-sample forecasting ability are mixed. A series of studies have assessed the predictive ability of different types of DSGE models. Christoffel et al. (2011) examine the out-of-sample predictive ability of the European Central Bank's New Area-Wide Model (NAWM), the DSGE model used to create projections of macroeconomic variables by the monetary authorities of the euro area. The results in Christoffel et al. (2011) indicate that this DSGE model, as compared to other alternative reduced-form specifications, provides good predictions for a set of 12 different macroeconomic variables. The predictive accuracy of DSGE models, however, does not necessarily remain stable over time. Del Negro et al. (2016) provide

☆ The authors would like to thank two anonymous referees for very helpful comments on an older version of the paper. Financial support from the Czech Science Foundation, Grants 17-14263S and 21-10562S, is gratefully acknowledged. This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic through the e-INFRA CZ (ID: 90140). Jesús Crespo Cuaresma gratefully acknowledges funding from IIASA, Austria and the National Member Organizations, Austria that support the institute. Niko Hauzenberger gratefully acknowledges financial support from the Jubiläumsfonds of the Oesterreichische Nationalbank (OeNB, grant no. 18718).

* Corresponding author at: Vienna University of Economics and Business, Austria.
E-mail address: jcrespo@wu.ac.at (J. Crespo Cuaresma).

https://doi.org/10.1016/j.ijforecast.2022.09.002
0169-2070/© 2022 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

evidence that forecasts produced using a Smets–Wouters type of DSGE model (Smets & Wouters, 2003, 2007) with financial frictions perform particularly well in periods of financial turmoil (in particular in the Great Recession), but that the predictive accuracy of the model tends to suffer in tranquil periods. The forecast quality of DSGE structures that include financial frictions has also been assessed by Kolasa and Rubaszek (2015), and improvements in forecast ability are reported in episodes of financial turmoil when housing market frictions are included in the model, although no systematic gains in predictive performance are found in more stable periods.

Another strand of the literature on macroeconomic forecasting has shown interest in analyzing predictive combinations based on a wide range of models, rather than focusing on a single specification, an idea that dates back to the work by Bates and Granger (1969). Amisano and Geweke (2017), for instance, find improvements in out-of-sample prediction errors for macroeconomic variables in the US by pooling forecasts from different macroeconomic models using Bayesian predictive distributions.

In this study, we evaluate the forecast ability of weighted combinations of six different DSGE models for GDP growth, inflation, and the interest rate in the euro area, making use of several prediction combination techniques. Our analysis expands the work by Wolters (2015), which assesses the forecast accuracy of four DSGE models for the US, as well as the potential predictive gains obtained by using combinations of these. We entertain six different DSGE specifications for the euro area and five forecast combination methods, both static and dynamic, and evaluate point forecasts as well as density predictions. Our set of prediction combination techniques contains some of the forecast pooling techniques entertained in existing studies for DSGE models (Wolters, 2015, for example), as well as more novel methods based on the optimization of weights, that can potentially be time-varying and evolve according to flexible laws of motion. In particular, we use static weights based on principles of Bayesian model averaging and prediction pools, and dynamic weights that build upon dynamic (latent) factor representations of the variables of interest.

The combination techniques employed in our analysis result in significantly different weighting schemes across models. While dynamic Bayesian model averaging and combinations based on dynamic factors lead to pooled forecasts which assign positive weights to all of the DSGE specifications, the technique based on prediction pools acts as a dynamic model selection tool, assigning weights close to zero to most individual model predictions over the out-of-sample period. The potential gains in predictive accuracy that can be exploited are specific to sub-periods, variables, and forecasting horizons, with no one-size-fits-all predictive combination strategy, ensuring systematic improvements in all situations.

The rest of this paper is organized as follows. Section 2 introduces the DSGE models used in the analysis, and Section 3 presents the predictive density combination methods. Section 4 shows the results of the out-of-sample forecasting exercise, and Section 5 concludes.

2. The battery of DSGE models

2.1. Individual DSGE models

For our empirical analysis, we use a battery of DSGE models for the euro area. Their specifications differ in size, complexity, and the particular features highlighted. Since the analysis is conducted on a set of three core macroeconomic variables (GDP growth, inflation, and the interest rate), we ensure that these three observable variables are common across all models. The sparsest model entertained is a basic three-equation New Keynesian model, which serves as a benchmark in terms of simplicity. The model presented in Cogley et al. (2010) also requires only three basic observable variables, but introduces two additional shocks and allows the inflation target to change over time. The specification by Christensen and Dib (2008) adds investment and money as additional observable variables. This group of models is extended with three more complex specifications that share the set of observable variables of the model by Smets and Wouters (2007): GDP growth, inflation, the interest rate, consumption growth, investment growth, real wage growth, and hours worked. The specification by Justiniano et al. (2011) contains the relative price of consumption to investment as the eighth observable variable, whereas Del Negro et al. (2015) add spread and inflation expectations as observable variables to the modeling framework and allow the shocks to be of a non-stationary nature. The group of DSGE specifications used spans model structures which differ in the mechanisms highlighted for the transmission of macroeconomic shocks. Tracking the predictive ability of such models over time can thus help us grasp changes in the relative importance of particular theoretical channels as determinants of macroeconomic dynamics in the euro area.

Table 1 lists the models entertained, together with their corresponding abbreviations (which are used in the description of the results of the analysis and in all subsequent figures and tables), and summarizes information about the number of observable variables, the number of exogenous shocks, and the main features of each model. The particular observable variables included in each one of the DSGE models are presented in Table 2.

2.2. Data

The models in Table 1 are estimated using quarterly data for the euro area in its 19-country composition. The database spans information from 1970Q3 to 2019Q1 and thus contains 195 quarterly observations. The core of the database is sourced from the Area Wide Model (AWM) presented in Fagan et al. (2005) and updated and extended by Brand and Toulemonde (2015). The original AWM database is updated using data from the European Central Bank or Eurostat since the 1990s and is extended by population and hours worked from the Total Economy Database and Eurostat. Data on monetary aggregates are obtained directly from the OECD. We use time series compiled by Gilchrist and Mojon (2018) for the interest rate spread variable. Inflation expectations are sourced from the European Central Bank's Survey of Professional

Table 1
Euro area DSGE models used in the forecasting exercise.

Reference	Name	Observables	Shocks	Features
3-equation basic NK model	NKModel	3	3	IS, PC, Taylor rule
Cogley et al. (2010)	CPS 2010	3	5	Inflation target can change over time
Christensen and Dib (2008)	CD 2008	5	5	Financial frictions as in Bernanke et al. (1999)
Smets and Wouters (2007)	SW 2007	7	7	Deterministic growth rate driven by labor-augmenting technological progress
Justiniano et al. (2011)	JPT 2011	8	8	Two investment shocks
Del Negro et al. (2015)	DNGS 2015	9	9	Smets and Wouters (2007) with financial frictions + spread and inflation expectations as observables

Table 2
DSGE models: Observable variables.

	CPS 2010 NKModel	CD 2008	SW 2007	JPT 2011	DNGS 2015
Output	✓	✓	✓	✓	✓
Inflation	✓	✓	✓	✓	✓
Interest rate	✓	✓	✓	✓	✓
Consumption			✓	✓	✓
Investment		✓	✓	✓	✓
Hours worked			✓	✓	✓
Wage			✓	✓	✓
Money supply (M1)		✓			
Relative investment price				✓	
Spread					✓
Inflation expectations					✓

Forecasters. The longest-term forecast available was selected, which spans four to five years ahead. Growth rates are calculated as quarter-on-quarter differences of logs, and the interest rate is calculated per quarter. Details on the sources of the different variables are provided in Appendix A.

The data transformations performed to the model variables correspond to those used in Smets and Wouters (2007). Real consumption, investment, and GDP are divided by population and transformed to growth rates. Hours worked are divided by population and logged. Inflation is defined as the growth rate of the GDP deflator. The nominal wage is deflated by the GDP deflator and transformed to growth rates. The interest rates are short-term market interest rates. The monetary aggregate M1 is deflated by the GDP deflator, divided by population, and transformed to growth rates. Finally, the relative price of investment is calculated as the investment deflator divided by the consumption deflator, and transformed to growth rates.

2.3. Detrending macroeconomic variables

The macroeconomic variables used in the estimation of DSGE specifications are often highly persistent and need to be detrended using methods that are consistent with the theoretical framework used in the model. For some existing models, the authors specify the particular filter employed to detrend the variables, while in other cases, these details are not specified (see, e.g., Gorodnichenko & Ng, 2010). Delle Chiaie (2009) investigates the effects of detrending observable variables with the Hodrick–Prescott (HP) filter and a linear trend in the model by Smets and Wouters (2003), and finds that structural parameter estimates are rather sensitive to the choice of a particular filtering method. Consequently, forecasting performance may be significantly affected by the choice of a detrending approach.

The original contributions on which we base our individual specifications use different detrending methods for the macroeconomic variables. While Christensen and Dib (2008) use the HP filters, Smets and Wouters (2007)—and models that build upon a similar structure—introduce some of the observable variables in first differences when estimating the parameters of the DSGE specification. Gorodnichenko and Ng (2010) offer a perspective of detrending approaches commonly used in a broader literature by compiling the detrending methods employed in 21 different models. The list of data filters used in various DSGE models shows a predominance of linear detrending, HP filtering, and first difference transformations. Our analysis employs several approaches used in the literature, while keeping the detrending method identical across all models considered. By doing so, we aim to separate the influence on forecasting performance of core model features, such as financial frictions or flexible inflation targets, from that of the trend formulation. In our baseline detrending specification, we use the data for GDP (and its sub-components) in first differences. Time series which present higher persistence are filtered using one-sided HP filters.[1]

[1] Alternatively, we also assess the forecasting performance of our models employing the filtering strategy proposed by Del Negro et al. (2015), Justiniano et al. (2011), and Smets and Wouters (2007) and find evidence that our baseline detrending approach leads to superior forecasting performance in the majority of cases (see Table C.4 in Appendix C). We also perform the analysis using different detrending approaches, such as using the (one-sided) HP filter for all data series, employing the regression-based data filter introduced in Hamilton (2018), and demeaning the times series in the models. The results for these alternative detrending specifications can be found in Appendix C.

2.4. Estimation and predictive densities

Each one of the models employed in the forecasting exercise is estimated recursively using Bayesian methods, starting with a sample size composed by 78 observations (corresponding to the time frame from 1970Q3–1989Q4) and adding one quarter at a time to the sample up to a maximum of 195 observations (corresponding to the full sample, which spans the period from 1970Q3–2019Q1). Additionally, we perform the forecasting exercise of estimating the models with a rolling window of 60 observations. The models are estimated using a minimum of a half million Metropolis–Hastings replications in two chains for the NKModel, one million replications for CD 2008 and CPS 2010, two million replications for JPT 2011 and SW 2007, and a minimum of four million replications for the DNGS 2015 model. To ensure convergence of the Markov chain, the checks in Brooks and Gelman (1998) are performed and, if these fail, the number of replications is increased until convergence of the posterior distributions is achieved. We use a Monte Carlo-based optimization routine to ensure that the optimal acceptance ratio of the Metropolis–Hasting algorithm is reached, and we discard 90% of the replications as burn-ins.[2]

Forecasts are computed using 10,000 draws from the posterior distribution for every estimated model and each in-sample period. In each instance, we calculate one- to four-step-ahead out-of-sample forecasts of GDP growth, inflation, and the interest rate, which correspond to periods ranging from 1990Q1 to 2018Q4. The analysis of forecasts is conducted after imposing back the trend of the observable variables, so as to ensure comparability across detrending approaches.

3. Predictive combinations of DSGE models

In this section, we outline the forecast combination methods employed to average the predictions of our set of models. Each DSGE model typically includes a different set of observables, targets a specific feature of the economy, and thus provides its own characterization of the economy by imposing different (structural) dynamics on the macroeconomic variables of interest. Some small-scale DSGE models abstract from the interaction between developments in the real economy, the labor market, and the financial sectors, while others include features and mechanisms related to these linkages. We concentrate exclusively on one- and four-step-ahead predictive densities of GDP growth, inflation, and the interest rate, which are common to all the specifications used. We assess and combine the joint predictive density of these three variables, as well as their corresponding marginal predictive densities.

In the following, we illustrate the methods we use to combine predictive densities by focusing on a scalar time series y_{t+1} and the one-step-ahead horizon. With only minor adjustments, these techniques work analogously for the joint predictive densities and for the multi-step-ahead horizon. In our analysis, we therefore consider predictive densities for y_{t+1}, which are available from K different DSGE models. Each DSGE model, indexed by $j = 1, \ldots, K$, incorporates information up to time t to generate a predictive density $p_j(y_{t+1} | \mathcal{I}_j(t))$ for period $t+1$. The information set $\mathcal{I}_j(t)$ typically consists of the target variable $\boldsymbol{y}_{1:t} = (y_1, \ldots, y_t)'$, as well as of the information up to time t provided by additional variables specific to that particular DSGE model. We aim to combine the K predictive densities $\{p_j(y_{t+1} | \mathcal{I}_j(t))\}_{j=1}^{K}$ using a $K \times 1$ weights vector $\boldsymbol{\omega}_{t+1} = (\omega_{1;t+1}, \ldots, \omega_{K;t+1})'$ that is specific to the one-step-ahead forecast horizon and potentially time-varying. The combined predictive density for y_{t+1} is then given by

$$p(y_{t+1} | \mathcal{I}_1(t), \ldots, \mathcal{I}_K(t)) = \sum_{j=1}^{K} \omega_{j;t+1} p_j(y_{t+1} | \mathcal{I}_j(t)). \tag{1}$$

Eq. (1) directly relates to the Bayesian predictive synthesis of McAlinn et al. (2019), McAlinn and West (2019), where $\boldsymbol{\omega}_{t+1}$ is described as a dynamic synthesis function.[3] This synthesis function can incorporate different objectives based on policy targets and historical performance up to period t, and nests traditional approaches to forecast combination, such as prediction pools (Geweke & Amisano, 2011; Hall & Mitchell, 2007) and Bayesian dynamic model averaging (Koop & Korobilis, 2012, 2013; Raftery et al., 2010). We start by discussing a simple static weighting scheme implying $\boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}$, and then turn to more general approaches based on using dynamic weights for the predictive densities.

Equal static weights

An obvious starting point to combine predictions from different DSGE models, which provides a benchmark to evaluate different weighting schemes, is to use

$$\omega_{1;t+1} = \cdots = \omega_{K;t+1} = 1/K.$$

Since $\omega_{j;t+1} > 0$ and $\sum_{j=1}^{K} \omega_{j;t+1} = 1$, the combination of predictive densities also constitutes a predictive density (Geweke & Amisano, 2011; Hall & Mitchell, 2007). This agnostic approach neglects the fact that different models might not be equally suitable for prediction at different time periods, and does not provide updates of the corresponding weights as information is gained about the differential predictive ability of model specifications. An equal weighting scheme is commonly found to be a good competitor in terms of out-of-sample forecasting accuracy, as it tends to hedge against large forecast errors of single specifications (see Timmermann, 2006).

Dynamic Bayesian model averaging

A natural choice of model weights can be achieved by pooling forecasts according to particular model selection criteria (for example, based on the predictive marginal likelihood or past forecast performance). For a given set

[2] All models are estimated using Dynare (Adjemian et al., 2011).

[3] Del Negro et al. (2016) and McAlinn and West (2019) provide a formal treatment of the decision problem concerning the choice of time-varying weights $\boldsymbol{\omega}_{t+1}$.

of priors over specifications, traditional Bayesian model averaging (BMA) approaches give models with a higher marginal likelihood more support while downweighting models with deficient predictive characteristics. Following Raftery et al. (2010) and Koop and Korobilis (2012, 2013), we consider posterior weights for individual specifications based on their (discounted) historical predictive likelihood, a procedure known as dynamic model averaging (DMA). According to this literature, DMA consists of a prediction equation

$$\pi_{j;t+1|t} = \frac{\pi_{j;t|t}}{\sum_{k=1}^{K} \pi_{k;t|t}};$$

and an updating equation

$$\pi_{j;t+1|t+1} = \frac{p_j\left(y_{t+1}^{(r)} \mid \mathcal{I}_j(t)\right) \times \pi_{j;t+1|t}}{\sum_{k=1}^{K} p_k\left(y_{t+1}^{(r)} \mid \mathcal{I}_k(t)\right) \times \pi_{k;t+1|t}}.$$

Here, $\boldsymbol{\pi}_{t+1|t} = (\pi_{1;t+1|t}, \ldots, \pi_{K;t+1|t})'$ denotes a $K \times 1$ vector of predictive weights at period $t+1$ based on historical forecast performance up to t, $\boldsymbol{\pi}_{t+1|t+1} = (\pi_{1;t+1|t+1}, \ldots, \pi_{K;t+1|t+1})'$ is a $K \times 1$ vector of updated weights, and $p_j\left(y_{t+1}^{(r)} \mid \mathcal{I}_j(t)\right)$ refers to the one-step-ahead predictive density for model j evaluated at the realized value $y_{t+1}^{(r)}$ (i.e., the predictive likelihood).[4] Moreover, a forgetting factor $\in (0; 1)$ discounts past predictive performance more heavily, while more recent predictive likelihoods receive more weight. In the empirical application, we set $= 0.95$, implying that the predictive likelihood four quarters (i.e., one year) in the past receives around 80% of the weight of the predictive likelihood of the most recent quarter.[5] The DMA algorithm, moreover, is easy to implement without the need for any simulation techniques.

Prediction pools

Recent approaches to forecast combination assess the set of model-specific forecasts as if it was a portfolio of predictions, which must be chosen optimally with respect to a particular loss function (see, inter alia, Geweke & Amisano, 2011, 2012; Hall & Mitchell, 2007; Pettenuzzo & Ravazzolo, 2016). Following Geweke and Amisano (2011), the loss function is defined as a function of historical log predictive scores, which gives rise to optimal weights after minimization. Similar to BMA and DMA methods, this approach ensures that forecasts from DSGE models with poor predictive abilities are downweighted, and those computed from specifications that predict more successfully receive higher weights. Information up to time t is available in order to choose the predictive weight $\boldsymbol{\pi}_{t+1|t}$ optimally. The negative weighted historical log predictive scores/likelihoods are minimized with respect to the weights vector $\boldsymbol{\pi}$:

$$\boldsymbol{\pi}_{t+1|t} = \min_{\boldsymbol{\pi}} \left\{ -\sum_{=1}^{t} \; {}^{t-} \log\left(\sum_{j=1} \pi_j p_j\left(y_{+1}^{(r)} \mid \mathcal{I}_j(\;)\right)\right) \right\};$$

where again denotes a discount factor that serves the same purpose as in the DMA procedure by assigning increasing weight to the most recent predictive performance. We additionally impose the restriction that weights are non-negative and sum to one. Note that we use standard numerical optimization algorithms for prediction pools, which are therefore easy to implement and computationally fast.

Bayesian predictive synthesis with a dynamic factor model

As noted by Del Negro et al. (2016), the predictive ability of particular specifications may be affected by structural breaks in the parameters governing the dynamics of macroeconomic variables. Such changes in predictive power should be addressed when combining the K predictive densities over time, and thus the mapping from the forecasts of each model to the combined predictive density should be adjusted accordingly. Eq. (1) can be directly related to a dynamic factor model, as proposed by McAlinn and West (2019) in the context of dynamic Bayesian predictive synthesis (BPS) methods, by defining the synthesis function as

$$y_{t+1}^{(r)} = \boldsymbol{F}_{t+1}' \boldsymbol{\pi}_{t+1} + \;_{t+1}; \quad \;_{t+1} \sim N(0; \;);$$

where we define the latent factors $\boldsymbol{F}_{t+1} = (\hat{y}_{1;t+1}, \ldots, \hat{y}_{K;t+1})'$ with $\hat{y}_{j;t+1}$, for $j = 1, \ldots, K$, being a draw from the one-step-ahead predictive density $p_j(y_{t+1} \mid \mathcal{I}_j(t))$ of each model j for period $t+1$. Further, $\boldsymbol{\pi}_{t+1}$ refers to time-varying loadings, and the shock in the observation equation $\;_{t+1}$ is Gaussian with zero mean and variance . The latent loadings (or states), that relate the draws from the predictive distributions to the realized value $y_{t+1}^{(r)}$ evolve according to a random walk:

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t + \;_{t+1}; \quad \;_{t+1} \sim N(\mathbf{0}; \;);$$

where $\;_{t+1}$ refers to a $K \times 1$ vector of Gaussian state innovations, which are centered on zero and feature a $K \times K$ variance–covariance matrix . In contrast to equal weighting, DMA, and predictive pooling, the weights $\boldsymbol{\pi}_{t+1}$ are no longer necessarily non-negative and do not need to sum up to one. $\boldsymbol{\pi}_{t+1}$ are thus to be interpreted as (time-varying) calibration parameters relating draws from the predictive densities to the actual realization $y_{t+1}^{(r)}$. A further difference from other weighting schemes is that we consider a measurement error $\;_{t+1}$ in the observation equation that explicitly accounts for model incompleteness (see, e.g., Aastveit et al., 2018; Hoogerheide et al., 2010; McAlinn & West, 2019). Moreover, the latent weights $\boldsymbol{\pi}_{t+1}$ are allowed to be correlated among models via a full variance–covariance matrix , which not only determines the amount of time variation introduced in $\boldsymbol{\pi}_{t+1}$, but also takes into account the dependencies

[4] By construction, both $\pi_{j;t+1|t}$ and $\pi_{j;t+1|t+1}$, for $j = 1, \ldots, K$, are non-negative, and the elements in both $\boldsymbol{\pi}_{t+1|t}$ and $\boldsymbol{\pi}_{t+1|t+1}$ sum to one.

[5] This choice is consistent with Koop and Korobilis (2012, 2013), who suggest defining $\in [0.95; 1]$. If $= 1$, past predictive performance is not discounted and the weights are defined according to the (predictive) marginal likelihood.

between individual predictive specifications that share similar characteristics.

We use weakly informative priors, which are standard in the literature for dynamic factor models. This implies the use of a multivariate normal prior for $\boldsymbol{\ell}_0$, an inverse Gamma prior for , and an inverse Wishart prior for . We repeat this procedure for R draws from the predictive density and explicitly account for a potentially non-trivial form of the predictive densities of the DSGE models. To estimate the model we rely on standard Bayesian estimation techniques used for time-varying parameter models. In particular, we use a Gibbs sampler which iterates through these R draws. Conditional on all other quantities, we update the latent states $\boldsymbol{\ell}_{t+1}$ with a standard forward filtering backward sampling (FFBS) algorithm (Carter & Kohn, 1994; Frühwirth-Schnatter, 1994). In a next step, conditional on the time-varying calibration parameters, we independently draw the observation equation variance and the state equation variance–covariance matrix . All steps involve standard conditional posteriors (for details, see McAlinn & West, 2019). Moreover, by using the filtering step in the FFBS algorithm, we directly obtain the predictive weights $\boldsymbol{\ell}_{t+1|t}$, which are used to combine the most recent predictive densities when the realization is not yet available. The MCMC algorithm of the dynamic factor model is somewhat more computationally demanding than the approximate procedure of DMA and the numerical optimization used for the pooling approach. However, compared to sequential Monte Carlo techniques such as particle filters (see, e.g., Billio et al., 2013; Del Negro et al., 2016), the computational burden can still be considered light.

The DECO approach

In addition to the combination methods outlined above, we consider the dynamic predictive density combination (DECO) approach of Billio et al. (2013). Like BPS, DECO allows for the specification of time-varying weights that evolve according to a flexible law of motion, and accounts for model incompleteness:

$$y_{t+1}^{(r)} = \boldsymbol{F}_{t+1}' \boldsymbol{\ell}_{t+1} + \quad {}_{t+1}; \qquad {}_{t+1} \sim N(0; \quad); \tag{2}$$

with $\boldsymbol{\ell}_{t+1}$ relating draws from the predictive densities to the actual realization $y_{t+1}^{(r)}$ and considering a Gaussian-distributed measurement error ${}_{t+1}$.

The main difference from BPS lies in the state equation that governs the evolution of the weights $\boldsymbol{\ell}_{t+1}$ and thus the learning mechanism used in prediction. Instead of assuming that the weights evolve according to a multivariate random walk with a full variance–covariance matrix , a non-linear link function between the elements in $\boldsymbol{\ell}_{t+1}$ and K independent dynamic latent processes ${}_{t+1} = ({}_{1;t+1}; \ldots; {}_{K;t+1})'$ is introduced:

$$\ell_{j;t+1} = \mathbb{P}\frac{\exp(\quad {}_{j;t+1})}{\sum_{j=1}^{K} \exp(\quad {}_{j;t+1})}; \quad \text{for } j = 1; \ldots; K.$$

This logistic link function does not allow for the use of unconstrained calibration parameters via a synthesis function, as in BPS, since it restricts the elements in $\boldsymbol{\ell}_{t+1}$ to be non-negative and sum to one. These restrictions thus effectively result in a non-linear state-space model, where Eq. (2) can be interpreted as a dynamic location mixture with a fixed variance. In what follows, ${}_{t+1}$ encodes the learning mechanism and governs the weight dynamics. Each element in ${}_{t+1}$ evolves according to independent random walks:

$${}_{j;t+1} = \quad {}_{j;t} + \quad {}_{j;t+1}; \qquad {}_{j;t+1} \sim \mathcal{N}(0; \quad {}_{j}); \quad \text{for } j = 1; \ldots; K.$$

Here, ${}_{j;t+1}$ denotes element-specific state innovations with zero mean and variance ${}_{j}$. In DECO, the state innovation variances ${}_{j}$ encode the learning mechanism and depend on a scoring rule preselected by the researcher, a discount factor , and a number of past observations considered. For example, if the scoring rule indicates that the predictive performance of some particular model has deteriorated for the past realized values, the mechanism allows for the corresponding adjustment of the weights by increasing ${}_{j}$, thus introducing time variation in $\ell_{j;t+1}$. Sequential Monte Carlo techniques are commonly used for such a non-linear state-space model. For the empirical implementation of DECO, we specify the key learning hyperparameters according to the following standard setting: we use the Kullback–Leibler scoring rule, set the number of past realized values to $= 9$, and the discount factor $= 0.95$. The remaining parameters are estimated from the data. For the particle filter, moreover, we define 50 particles and use an additional smoothing factor of 0.01.[6]

4. Forecasting macroeconomic variables in the euro area using DSGE models

We start by quantitatively assessing the predictive ability differences across DSGE models, before moving to the analysis of the potential improvements in forecasting quality from combining the predictions of individual models, and of the dynamics of predictive weights over the out-of-sample period.

4.1. Overall forecast performance of individual DSGE models

The top panel of Table 3 presents the forecasting performance of individual DSGE models, which are estimated recursively over the out-of-sample period. We present the root mean squared forecast error (RMSE) ratios, as well as the average log predictive Bayes factors (LPBFs), defined as the difference in average log predictive scores (LPSs), for one-step-ahead and four-step-ahead predictions. For the RMSEs, Table 3 also shows the results of Diebold–Mariano tests of equal predictive performance (Diebold & Mariano, 1995), and for the LPSs, those of the Amisano–Giacomini tests (Amisano & Giacomini, 2007). In both cases, the equality of predictive ability is tested using the SW2007 model as the benchmark specification. The results of this predictive ability analysis based on rolling window estimation (instead of parameter estimates based

[6] An efficient algorithm for this approach is implemented in the *DeCo toolbox* in Matlab (see Casarin et al., 2015) for one-step-ahead forecasts.

Table 3
Forecasting performance of recursively estimated DSGE models and combinations of these models.

Target variable(s)	DSGE model					
	CD 2008	CPS 2010	DNGS 2015	JPT 2011	NKmodel	SW 2007
One step ahead						
Joint	1.246***	0.929**	1.069**	1.196***	**0.889****	0.320
	(−0.194)	(**0.340***)	(−0.063**)	(−0.452)	(0.159*)	(0.125)
GDP growth	1.277***	0.938	1.069**	1.206***	**0.885***	0.511
	(−0.236**)	(0.064*)	(−0.069)	(−0.224***)	(**0.102***)	(−0.777)
Inflation	1.033	**0.858*****	1.064	1.107**	0.896**	0.194
	(−0.013)	(**0.117***)	(0.036***)	(−0.089)	(0.019*)	(0.022)
Interest rate	1.122	**0.973***	1.088	1.291***	1.016	0.085
	(−0.011*)	(**0.138***)	(−0.017)	(−0.167***)	(0.012)	(0.819)
Four steps ahead						
Joint	1.115	0.981	1.017	1.176***	**0.941****	0.379
	(0.228)	(**0.419***)	(0.308***)	(−0.255***)	(0.418***)	(−1.219)
GDP growth	1.115	0.996	1.037*	1.148**	**0.963**	0.573
	(−0.162)	(−0.041)	(−0.043)	(−0.241***)	(−0.026)	(**−0.865**)
Inflation	1.099	0.863	0.938	1.375***	**0.775****	0.220
	(0.268***)	(**0.388***)	(0.268***)	(−0.006)	(0.331***)	(−0.396)
Interest rate	1.128	0.988	0.968	1.148***	**0.940**	0.234
	(0.051***)	(0.122***)	(0.089***)	(−0.038*)	(**0.162***)	(−0.101)
	Combination method					
	EQ	DMA	POOL	BPS	DECO	
One step ahead						
Joint	1.061*	**0.926**	0.936	0.993	0.941	
	(0.085)	(0.328***)	(**0.330***)	(0.164***)	(0.226***)	
GDP growth	1.074*	**0.928**	0.942	1.003	0.949	
	(−0.006)	(0.072***)	(**0.085***)	(−0.011)	(−0.145***)	
Inflation	0.958	0.877**	0.884**	0.929**	**0.872*****	
	(0.028)	(0.082)	(0.115***)	(0.109***)	(**0.307***)	
Interest rate	1.102	1.086	0.994	**0.925**	0.987	
	(0.017)	(0.102***)	(0.123***)	(0.118***)	(**0.158***)	
Four steps ahead						
Joint	1.070**	0.999	0.985	**0.926**		
	(0.321***)	(0.440***)	(0.448***)	(**0.673***)		
GDP growth	1.089**	1.015	1.005	**0.974**		
	(−0.020)	(**0.008**)	(−0.037)	(−0.013)		
Inflation	1.019	0.896	0.871	**0.831****		
	(0.234***)	(0.365***)	(0.386***)	(**0.396***)		
Interest rate	0.997	0.991	0.961	**0.667****		
	(0.108**)	(0.143***)	(0.150***)	(**0.409**)		

Notes: The table shows root mean squared errors (RMSEs), and average log predictive Bayes factors (LPBFs) in parentheses, relative to the SW 2007 model. Bold numbers indicate the best performing DSGE model as well as the best performing combination method that obtains the smallest RMSE ratio (largest LPBF). The SW 2007 column (highlighted in gray) shows the actual RMSEs and log predictive scores (LPSs) of our benchmark. Asterisks indicate statistical significance relative to the SW 2007 model at the 1% (***), 5% (**), and 10% (*) significance levels in terms of Diebold and Mariano (1995) tests for RMSEs and Amisano and Giacomini (2007) tests for LPSs.

on enlarging the in-sample period recursively) can be found in Appendix B, and the results based on alternative detrending methods are presented in Appendix C. The forecast error measures are presented for the joint vector of GDP growth, inflation, and the interest rate, as well as for these three variables individually.

We start by considering the overall forecasting ability for the group of macroeconomic variables, reflected in the characteristics of the joint predictive distribution. The results in the top panel of Table 3 for the full out-of-sample period indicate that the simple NKModel has particularly good predictive ability compared to other DSGE specifications with more complex model structures. In terms of the joint accuracy of point forecasts (i.e., for the full vector of variables) as measured by the average RMSEs, this specification outperforms all other DSGE models for both one-step-ahead and four-step-ahead predictions. Considering each variable individually, the quality of point predictions of the NKModel appears particularly high for four-step-ahead predictions.

The quality of point forecasts from the NKModel partially translates to good performance in density forecasting (as measured by the LPBFs) in both of the prediction horizons considered. The joint density predictions of the NKModel specification, however, appear less accurate than those of the CPS2010 model, which includes five structural shocks instead of the three of the NKModel. The focus of the CPS2010 specification on offering a structural modeling framework for inflation dynamics (based on the inclusion of changes in the inflation target in the model) is successful at improving out-of-sample density predictions for this variable compared to the rest of the DSGE

models entertained. Furthermore, the average predictive performance of the CPS2010 model for short-run density forecasts of the interest rate is also the best among the set of models considered.

The particularly good forecasting ability of models that include a small number of observable variables is broadly robust to the use of different detrending methods and to the use of parameter estimates based on a rolling sample instead of on recursive estimation (see Appendices B and C).

4.2. Overall forecast performance of predictive combinations

The comparison of results concerning predictive ability presented in Table 3 indicates that using forecast combinations can lead to improvements in average predictive ability over the full out-of-sample period. The best individual models in terms of forecasting ability at short horizons outperform all of the combination methods for GDP growth and jointly for all three variables. Concentrating on point prediction performance and both analyzed horizons, individual DSGE models predict GDP and inflation better than any combination scheme considered. The combinations, on the other hand, outperform individual DSGE specifications at predicting the interest rate. Combining predictions of DSGE models also delivers better results for density forecasting inflation, and yields the best results when evaluating the longer horizon of joint predictive performance.

Since the forecasting ability results of single DSGE specifications and their combinations for the full sample may be driven by differences in out-of-sample predictive quality in sub-periods of the out-of-sample interval chosen, a more detailed analysis of the dynamics of the weights that combination methods assign to different DSGE models appears necessary. In the following sub-section, we analyze the dynamics of the predictive weights for the different averaging methods entertained, thus moving beyond average forecast quality and turning to the assessment of changes in predictive accuracy over time.

4.3. The dynamics of predictive weights

We start by assessing the dynamics in the relative predictive ability of DSGE models by studying the evolution of predictive weights along the hold-out sample for our three target variables: GDP growth, inflation, and the interest rate. For each observable variable, we combine the predictions from DSGE models using statistics based on marginal predictive densities rather than on the joint predictive density of all target variables. One key advantage of this approach is that the weights used to combine predictive densities are thus specific to each variable and reflect changes in the relative forecasting ability of each DSGE specification for that particular phenomenon. We calibrate the weights for each forecast combination scheme with at least eight quarters (1990Q1 to 1991Q4) for the first period of our hold-out sample (1992Q1) and employ $= 0.95$.

Figs. 1 and 2 show the weights obtained for each model and target variable in the hold-out sample period for one-step-ahead (Fig. 1) and four-step-ahead forecasts (Fig. 2). The weighting schemes across forecast horizons are relatively similar, indicating a certain degree of stability of the predictive power of DSGE models across forecast horizons. In spite of the fact that the loss functions in the DMA and prediction pool methods are both based on log predictive scores, we observe substantial differences in the magnitude of the weights obtained for these two approaches. The weights in the prediction pool approach typically suggest a dynamic model selection scheme where single models tend to receive a weight close to one in a given period of time, while DMA usually assigns positive weights to forecasts from all different DSGE models. For the combination approach based on Bayesian predictive synthesis, weights (corresponding to factor loadings) are positive and relatively similar across models for the majority of periods. However, during the financial crisis, individual negative factor loadings can be observed, implying a reversal of the sign of the prediction of the respective DSGE model in the combined forecast for these quarters.

Focusing on one-step-ahead weights, the first row of panels in Fig. 1 shows the results for the different combination techniques for GDP growth. For DMA, we observe that CPS2010 and NKModel tend to dominate in terms of predictive ability prior to the financial crisis. In the subsequent years, and in particular after the debt crisis in the euro area, the relevance of CPS2010 within the group of combined predictions decreases in favor of SW2007. For prediction pooling, the distribution of weights shows the importance of predictions from CPS2010 and NKModel for the combined forecast in particular periods, with SW2007 gaining importance only in the aftermath of the debt crisis. Both the DMA and DECO combination schemes give high weights to predictions from CPS2010 and NKModel, and the weights from DECO reflect the importance of forecasts from DNGS2015 until the mid-2000s. The distribution of weights implied by Bayesian predictive synthesis is much more uniform and stable over time.

The second row of panels in Fig. 1 depicts the dynamics of weighting schemes for inflation as a target variable for one-step-ahead forecasts. Using DMA, the highest weights are assigned to CPS2010 and DNGS2015, with the latter gaining importance during the financial crisis. Both of these models are designed with a focus on tracking inflation dynamics: CPS2010 features a time-varying inflation target, and DNGS2015 includes inflation expectations, operationalized by making use of data from the Survey of Professional Forecasters. With prediction pools, a qualitatively similar scheme appears, with weights close to unity alternating between these two DSGE models, and predictions from DNGS2015 being particularly important during the financial crisis years. Bayesian predictive synthesis and DECO assign practically identical stable weights across models for the full period.

For interest rate predictions, the resulting weighting schemes are presented in the third row of panels in Fig. 1. In general, for the interest rate we observe a more persistent pattern in the weighting scheme, similar to that found for inflation. The DMA method leads to large and stable weights for CPS2010 throughout the hold-out

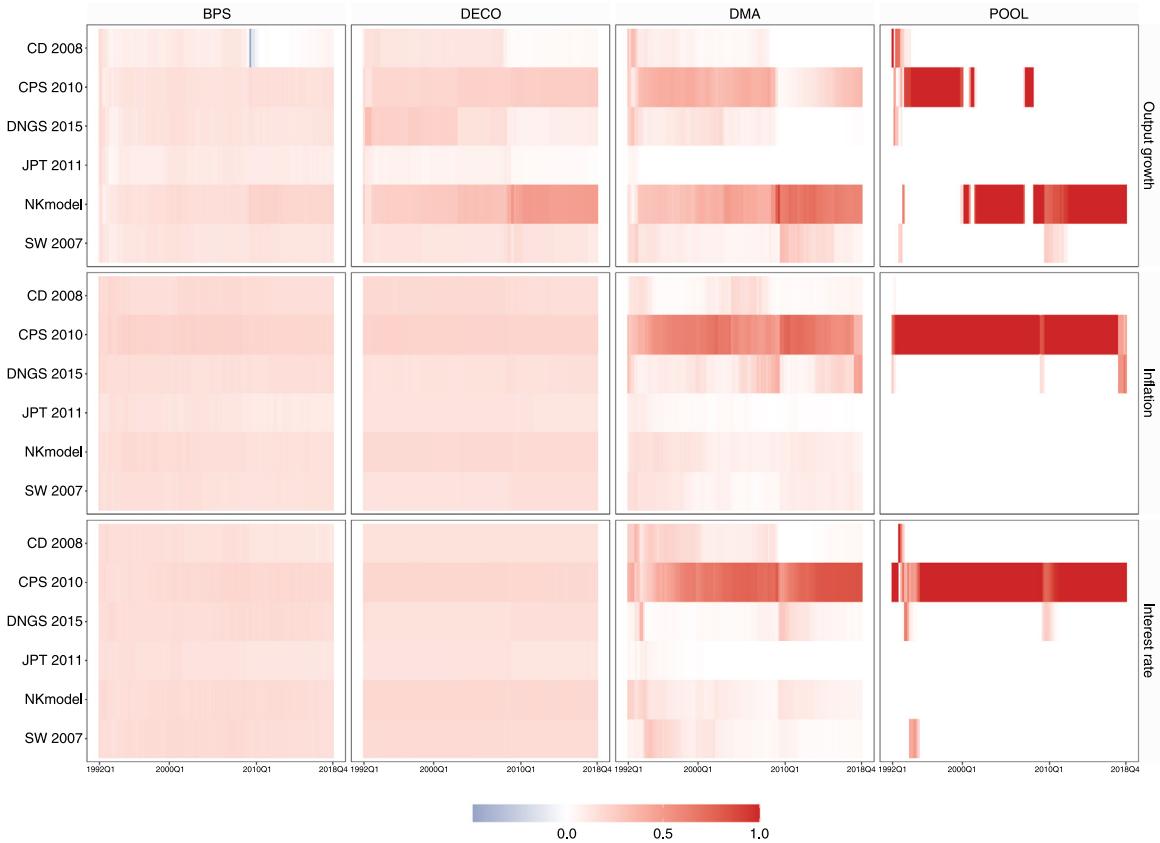

Fig. 1. Evolution of model weights over the hold-out sample for one-step-ahead predictions. *Notes:* The figure shows four different weighting schemes for the three target variables: GDP growth, inflation, and the interest rate. For BPS and DECO we use the posterior mean as a point estimate.

sample, with the exception of the period corresponding to the financial crisis, when DNGS2015 and NKModel receive relatively larger weights. The results from prediction pools are qualitatively similar, with forecasts from CPS2010 receiving weights close to unity throughout the period, except for in the mid-1990s and during the financial crisis, where predictions from SW2007 and DNGS2015 play a small role. As in the case of inflation, for interest rates, Bayesian predictive synthesis and DECO assign stable and similar weights to the individual model predictions throughout the hold-out sample.

For four-step-ahead forecasts of GDP growth, Fig. 2 shows a partly similar evolution of the weights for DMA combinations, but with weights that are more spread across DSGE specifications, especially before the financial crisis. In contrast to one-step-ahead predictions, for the longer horizon, the forecasts of GDP growth from SW2007 gain importance during the euro area debt crisis period, and weights in the last part of our hold-out sample are more uniformly spread across DSGE specifications. For output, the combination chosen by prediction pooling leads to a more erratic weighting scheme prior to the financial crisis as compared to one-step-ahead predictions. Output growth forecasts from CD2008 gain relevance right before the financial crisis, as do those from NKModel and SW2007 in the aftermath of the debt crisis in the euro area. The weights from the combination method based on Bayesian predictive synthesis for four-step-ahead forecasts roughly resemble those found for one-step-ahead predictions.

The evolution of weighting schemes along the hold-out sample for inflation predictions at the four-step-ahead horizon is relatively similar to that for the one-step-ahead predictions. The pooling combination scheme selects the CPS2010 model for almost the whole time period under study, as in the case of the shorter prediction horizon. More notable differences across prediction horizons can be found for DMA combinations. For the longer prediction horizon, the JPT2011 and SW2007 models are assigned almost zero weight, while DNGS2015 receives higher weight in the aftermath of the debt crisis in the euro area. The particular characteristics of the DNGS2015 model, which includes financial frictions and aims to explain the dynamics of output and inflation after financial shocks, make it conceptually adequate for predictions in the environment of debt distress. The Bayesian predictive synthesis combination method results in roughly uniformly distributed weights across models.

Finally, the results for interest rate predictions at the four-step-ahead horizon, presented in the last row of Fig. 2, differ strongly from those obtained for one-step-ahead forecasts. The predictions of the CPS2010 model, which obtained the highest weights using DMA and prediction pools for the shorter-term horizon, now receive

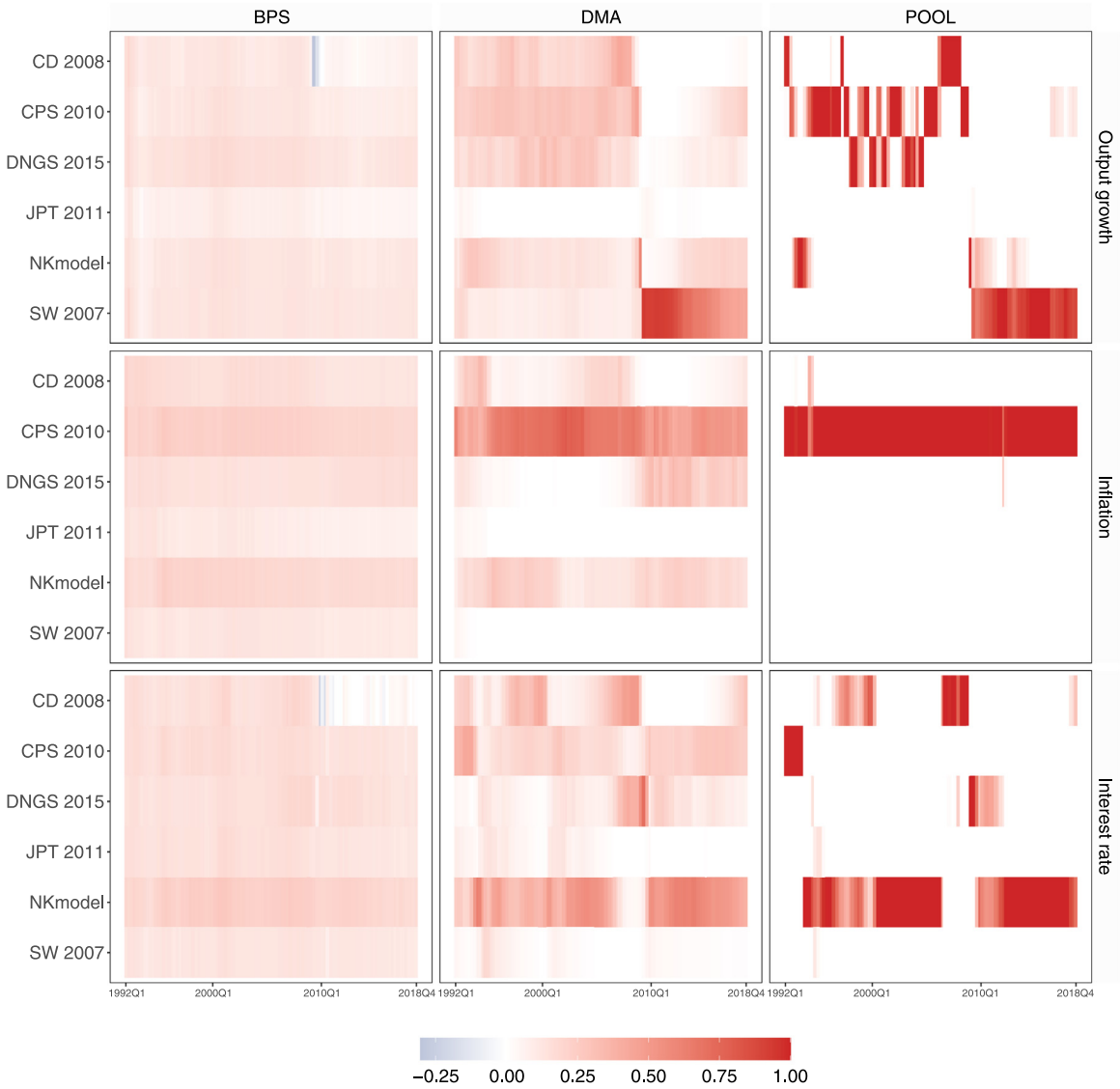

Fig. 2. Evolution of model weights over the hold-out sample for four-step-ahead predictions. *Notes:* The figure shows three different weighting schemes for the three target variables: GDP growth, inflation, and the interest rate. For BPS we use the posterior mean as a point estimate. Note that DECO is only used for the one-step-ahead horizon.

low weights over the hold-out sample and are replaced by the NKModel for the majority of the hold-out period, with the weights for CD2008 and DNGS2015 being prominent during the outbreak of the financial crisis.

The results of the analysis of the evolution of weight estimates for combinations of DSGE model predictions illustrate the stark differences in weights across forecast pooling methods and over time. The fact that the combination method based on prediction pools acts as a dynamic model-selection device contrasts with the weighting schemes resulting from the other approaches entertained in the exercise, which tend to lead to composite predictions with positive weights for all specifications. The relative predictive performance of these combination approaches along the hold-out sample, as well as that of individual model forecasts, is explored in more detail in the next section.[7]

4.4. Predictive ability of individual specifications and forecast combinations: Variation over time

In this section, we examine the variation over time of the predictive performance of the individual DSGE models and the forecast combinations. We concentrate on the

[7] The evolution of predictive weights across methods and over time for rolling samples can be found in Appendix B.

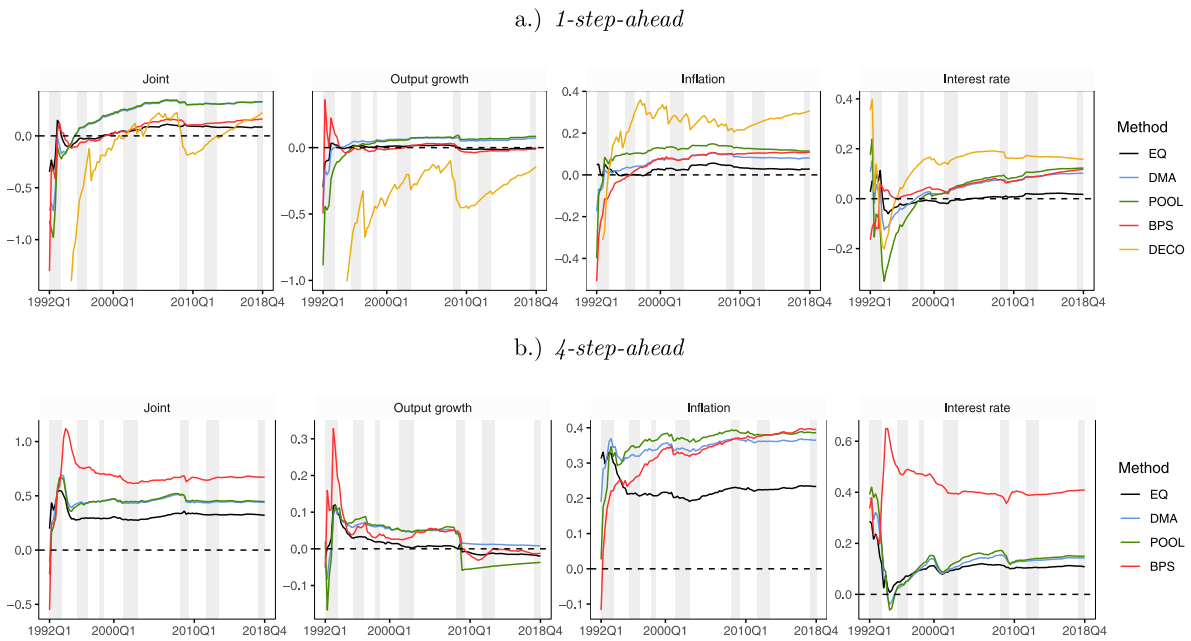

Fig. 3. Evolution of average log predictive Bayes factors (LPBFs) relative to the SW 2007 model. Combination methods. *Notes:* The gray shaded areas indicate OECD recessions for the euro area. Note that DECO is only used for the one-step-ahead horizon.

analysis of the evolution of log predictive Bayes factors, as a measure for the marginal likelihood, over the hold-out sample.

Fig. 3 presents the predictive performance of forecasts based on the different weighting schemes across variables and forecast horizons by means of log predictive Bayes factors relative to the SW2007 model. In panel a) of Fig. 3, the results for one-step-ahead forecasts are shown. The overall evolution of the predictive ability of forecast combination methods at this prediction horizon presents similar dynamics across most of the approaches, with improvements in predictive ability over the hold-out sample and a relatively stable forecasting performance at the end of the out-of-sample period. A notable exception is the DECO scheme, especially for output growth and inflation. Practically all forecast combination methods tend to perform poorly at the very beginning of our hold-out sample compared to the SW2007 benchmark, a feature that is likely related to the imprecise estimation of weights.[8]

Considering the joint set of macroeconomic variables of interest as a whole, the predictive ability of prediction pooling and DMA tends to be similar and to dominate all other combination methods after the mid-1990s, a result which is mostly driven by their ability to provide precise predictions of GDP growth. Combinations of forecasts based on the DECO method, on the other hand, dominate the other combination alternatives when predicting inflation and interest rates after the mid-1990s. In contrast to the results obtained for the shorter-term horizon, the Bayesian predictive synthesis method of forecast averaging systematically outperforms the other predictive combinations for the joint group of observable macroeconomic variables after the mid-1990s at the longer horizon. The predictive quality shown by this method is fueled by its performance at predicting interest rates in the longer term, while in the other two variables, the forecast error appears comparable to that of other combination methods.

In Fig. 4 we present the log predictive Bayes factors of individual specifications over the hold-out period with respect to the benchmark model, SW2007. A comparison across DSGE models reveals a systematically good relative predictive performance of the CPS2010 model (in particular after the mid-1990s) that extends to all three variables and to both forecasting horizons. In addition, a worsening in forecast ability of some specifications with respect to the SW2007 benchmark during the financial crisis and in its aftermath can be observed for many of the individual DSGE specifications. This is particularly the case for CD2008 at both horizons, but the loss of predictive quality also takes place in other specifications and is asymmetric across macroeconomic variables, with GDP growth forecasts being the most affected. The loss of predictive power triggered by the financial crisis is in many cases persistent, and relative predictive scores (as measured by the log predictive Bayes factor) do not always reach the level they had prior to the crisis. An interesting exception to this stylized fact is the inflation

[8] We also perform the exercise based on rolling samples instead of a recursive reestimation scheme, and the results are presented in Appendix B. The relative forecasting ability of individual models does not change qualitatively, while the performance of combination schemes with respect to the SW2007 benchmark tends to worsen, thus lending support to this conclusion.

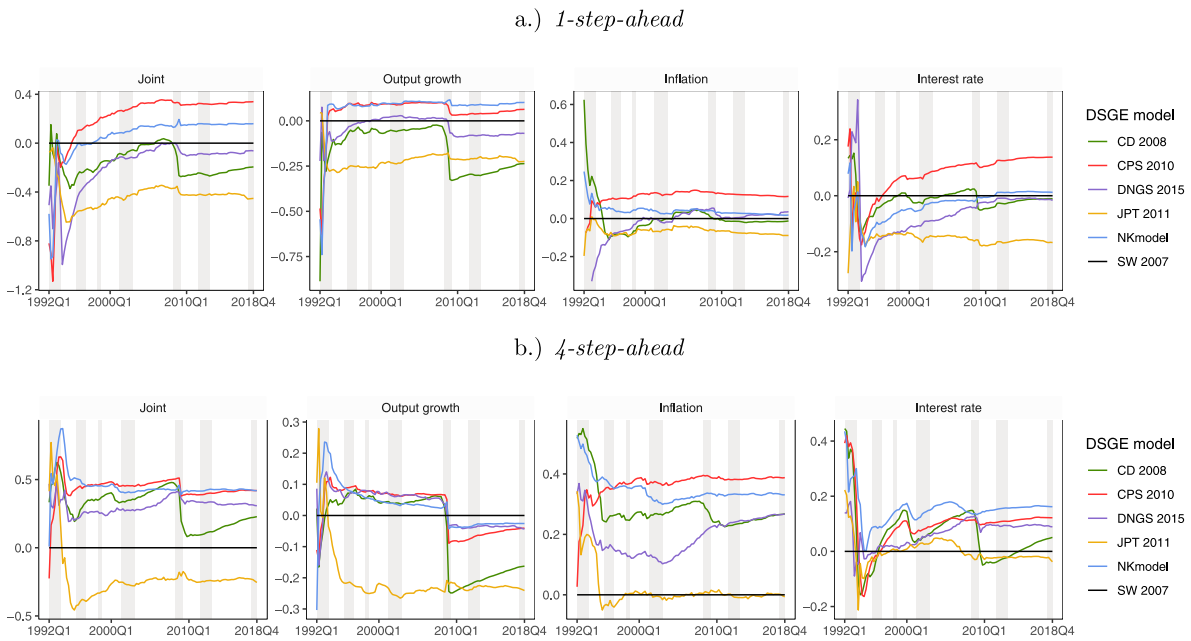

Fig. 4. Evolution of average log predictive Bayes factors (LPBFs) relative to the SW 2007 model. DSGE models. *Notes:* The gray shaded areas indicate OECD recessions for the euro area.

predictions from the DNGS2015 model, whose specification incorporates a more sophisticated assessment of inflation expectations than the rest of the DSGE models used, and whose predictive ability for this variable improves in the crisis period.

A comparison of the predictive ability of forecast combinations and individual DSGE models over the hold-out period reveals that in some periods and for particular variables, weighted averages of forecasts achieve higher and less volatile log predictive Bayesian factors. However, the results show that it is not possible to find a one-size-fits-all method to combine predictions from DSGE models that would provide systematically superior predictions for all variables under scrutiny and over the full period studied. The difficulty in finding such a forecast averaging method for our sample is related to the particular characteristics of the economic area being studied. The existence of cross-country heterogeneity in shock transmission mechanisms and macroeconomic outcomes across euro area economies, in particular since the onset of the sovereign bond crisis, is widely documented in the literature (see Burriel & Galesi, 2018; Holton & d'Acri, 2018, just to name two recent examples). The difference in shock propagation between countries in the euro area aggregate poses particular challenges in terms of how they can be accommodated in DSGE specifications such as those entertained in our analysis.

5. Conclusions

The results of our analysis show that combining forecasts from DSGE models does not systematically lead to improvements in predictive ability for macroeconomic variables for the euro area over the full period under scrutiny, which spans the last three decades. For some variables and periods, predictive weighting schemes are able to reach superior forecasting performance over individual DSGE specifications. In particular, the gains in the predictive ability of forecast combinations of DSGE models are larger in the last part of our sample.

The weighting schemes implied by the combination methods employed are fundamentally different across techniques. Weighting based on prediction pools tends to lead to forecasts based on dynamic model selection, assigning zero weights to many individual model predictions over the out-of-sample period. DMA and weighting based on dynamic factors, on the other hand, results in combined forecasts with positive weights for practically all of the DSGE specifications. The forecasting performance of individual DSGE models and combinations thereof systematically worsens during the financial crisis with respect to the benchmark, although the loss of predictive power and the volatility of forecast errors appear larger in individual specifications as compared to predictive combinations.

The results of our analysis may be significantly affected by the focus on the euro area economy, which is characterized by differences in the propagation of macroeconomic shocks across the countries that compose it. The suite of DSGE models employed in our forecasting exercise does not contain any specification that explicitly addresses the differential structural characteristics of the euro area. In this context, the results of our analysis should be considered very conservative estimates of the potential of predictive combination methods combined with forecasts from DSGE models. Refining the theoretical