Density Forecast Combinations: The Real-Time Dimension

Peter McAdam[#] and Anders Warne[*]

October 23, 2023

Abstract: Euro area real-time density forecasts from three DSGE and three BVAR models are compared with six combination methods over the sample 2001Q1–2019Q4. The terms information and observation lag are introduced to distinguish time shifts between data vintages and actuals used to compute model weights and compare the forecast, respectively. Bounds for finite mixture combinations are presented, allowing for benchmarking them given the models. Empirically, combinations with limited weight variation often improve upon the individual models for the output and the joint forecasts with inflation. This reflects over-confident BVAR forecasts before the Great Recession. For inflation, a BVAR model typically performs best.

Keywords: Bayesian inference, euro area, forecast comparisons, model averaging, prediction pools, predictive likelihood.

JEL Classification Numbers: C11, C32, C52, C53, E37.

1. Introduction

The benefits of combining density forecasts from different models or forecasters have long been recognized across many academic fields, such as management science, meteorology and statistics. Density forecast combinations have also attracted a growing interest among economists and policy makers. Not only do combinations provide a way to guard against model uncertainty, it is furthermore a means to improve forecast accuracy; see Timmermann (2006) and Aastveit et al. (2019). The improvement in forecast accuracy can, for instance, arise from individual models being over- or under-confident in the sense of delivering predictive densities that are too narrow or too wide and thereby not *well-calibrated*; see, e.g., Dawid (1984) and Diebold et al. (1998).

Notwithstanding this positive consensus on forecast combinations, there is less empirical agreement on the performance and robustness of different combination schemes. Different methods can generate different outcomes and reflect different philosophies; see Amisano and Geweke (2017). For

Acknowledgements: We thank two anonymous referees and editor Massimiliano Marcellino. We have also benefitted from discussions with and suggestions by Gianni Amisano, Romain Aumond, Dean Croushore, Szabolcs Deak, Mátyás Farkas, Domenico Giannone, Gary Koop, Michele Lenza, Giorgio Primiceri, Bernd Schwaab and Allan Timmermann, as well as from participants at seminars at the European Central Bank and at the University of Kent, Canterbury. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Central Bank or the Eurosystem, or the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

Corresponding author: Economic Research Department, Federal Reserve Bank of Kansas City, 1 Memorial Drive, Kansas City, Missouri 64198, USA; peter.mcadam@kc.frb.org; Phone: +1-816 585 0118.

* Forecasting and Policy Modelling Division of DG-Economics, European Central Bank, 60640 Frankfurt am Main, Germany; anders.warne@ecb.europa.eu; Phone: +49-69 1344 8737.

instance, the well-known method of Bayesian model averaging is predicated on the assumption of a complete model space, while optimal prediction pools, suggested by Hall and Mitchell (2007), make no such assumption: all models in the pool may be false, but nonetheless useful. Straddling these extremes is the enduring puzzle that naïve schemes, such as equal model weights, often outperform more sophisticated alternatives. Equal weighting, though, precludes the possibility of adaption to particular episodes of model improvements. If the forecast horizon contains some dramatic event or particular constellation of shocks, this may be costly. On the other hand, schemes that yield volatile model weights may undermine the practical case for combination methods.

Against this background, our paper makes three principal contributions. First, like forecasting itself, we believe gains from combinations matter most in real time. This is because real-time data constitutes the most realistic and policy-relevant testing ground. Many studies have assessed density forecasting with different competing models using real time data, e.g., Jore et al. (2010), Clark (2011), Groen et al. (2013), Mazzi et al. (2014); see also Chauvet and Potter (2013) and Clements (2017) for a discussion of real-time forecasting issues.[1] However, what is missing from this literature is a formal recognition that the use of real time data has implications for the performance and analysis of combination schemes.

Apart from fixed-weight combinations, weights are usually computed using information about each model's (or forecaster's) past predictive performance. In a real-time context, however, model weighting emerges when outcomes are imperfectly known: measured values of the predicted variables for period t are, by construction, not observed until later. This implies that the predictive measures, such as the predictive likelihood, should be suitably lagged when computing the incremental weights.

To that end we define the following terms: the *observation lag* is the time difference between the date of the variable and the vintage its actual or "true" value is taken from; while the *information lag* is the time difference between the date of the vintage and the date of the last data point of the predicted variables which is used for computing the model weights when forecasting with this vintage. The former concept concerns the data used for the performance measure of density forecast combinations, while the latter is related to the information used for computing combination weights. Note that the information lag comes on top of the forecast horizon so that the sum of the two make up the delay before historical density forecasts can be used for model weighting. This

[1] Other papers of note which separately consider real-time forecasting or combinations exercises include Edge et al. (2010) and Rossi and Sekhposyan (2014).

requires us to re-specify the density combination method to accommodate these features. We then demonstrate that the assumed length of the information lag matters for the attainable gains from forecast combinations and the ranking of the methods over time.

In this context, it is also worth emphasizing that we consider forecasts at various horizons in our real-time analysis: backcasts, nowcasts and up to eight-quarter-ahead forecasts. This matters because models' predictive performance can be highly horizon specific. In that respect, the Great Recession episode is telling: all models incur large forecast errors, but some "recover" better than others over different horizons and for different reasons. This has implications for the gains obtainable from combinations in general as well as the specific performance of different combination schemes. A standard one-step-ahead density forecast would suppress these issues. In addition, the impact of the assumed information lag on the performance of combination methods may vary with the horizon, where one may a priori suppose that its influence is greater at short than at long forecast horizons.

Second, we introduce *upper and lower bounds* for finite mixtures of the model density forecasts, allowing us to benchmark such combination methods not only with respect to the available models but against the best and the worst cases given the models involved. The bounds are computed from the density forecasts of the models using the actual values, implying that they are *ex post* bounds. They tells us what can be achieved by finite mixture combination methods *for the given set of models*. If the best-performing model is far from the upper bound, then there is a large room for improvement. If near the upper bound, then finite mixture combination methods are unlikely to be useful and either additional models may be considered before combinations are used or compound distribution based methods may be contemplated. This constitutes, we believe, an important practical diagnostic aid, which has not been considered in the literature.

Finally, we contribute to the literature on combinations in a *euro area context*. Relative to that of the US, evidence of real-time density forecasting on euro area data remains scant, despite its similar economic weight, and corresponding evidence on combinations is scanter still. This is also important since the euro area real time data has distinct properties and constraints in terms of sample size and number of variables. Notable papers using the euro area real time database include Mazzi et al. (2014), Smets et al. (2014), Berg and Henzel (2015), Jarociński and Lenza (2018), Bańbura et al. (2021) and Warne (2022).

The paper is organized as follows. Section 2 discusses probabilistic forecasting with a focus on combination methods and the real-time dimension. Section 3 overviews the models used: three

DSGE models that are variants of the canonical Smets and Wouters model (McAdam and Warne, 2019), as well as three Bayesian vector autoregressions (BVARs), embodying standard and recently developed priors. In Section 4, the forecast performance of the six models is presented for the sample 2001Q1–2019Q4, a period which constitutes an especially challenging laboratory: it spans a period of relatively calm macroeconomic conditions, undone by the Great Recession. Since we make use of annual revisions for the actual values of the forecasted variables vintages until 2020Q4 are utilized. In Section 5 we define the upper and lower bounds for the combination methods based on the models at hand, compare the predictive performance of the models relative to the different combination schemes and study the combination weights. The sensitivity of the results to the information lag assumption is also examined. Finally, Section 6 summarizes the main findings, while additional material is in the Online Appendix.

2. Density Forecast Combinations in Real Time

Scoring rules are widely used to compare the quality of probabilistic forecasts by attaching a numerical value based on the predictive distribution and an event or value that materializes; see Gneiting and Raftery (2007) for a survey on scoring rules, and Gneiting and Katzfuss (2014) for a review on probabilistic forecasting. A scoring rule is said to be *proper* if a forecaster who maximizes the expected score provides his or her true subjective distribution, and it is said to be *local* if the rule only depends on the predictive density at the realized value of the predicted variables. A well-known scoring rule is the log predictive score and it is the only proper local scoring rule; see Bernardo (1979). The empirical analysis below relies on this rule as it is the most commonly applied scoring rule in practice.[2]

2.1. Log Predictive Score

Suppose there are M models to compare in a density forecast exercise. Let $p^{(i)}_{t+h|t} = p(x^{(a)}_{t+h}|\mathcal{I}^{(i)}_t, A_i)$ denote the predictive likelihood conditional on the assumptions of model i, A_i, and the information set of model i, $\mathcal{I}^{(i)}_t$. The predictive likelihood is given by the predictive density evaluated at the actual or observed value of the vector of random variables x, realized at time $t+h$ and denoted by $x^{(a)}_{t+h}$, with the integer h being the forecast horizon. The log predictive score of model i for

[2] If we remove the requirement of using the predictive density for the scoring rule and allow for the predictive cumulative distribution, then additional proper and strictly proper scoring rules exist. Examples include the univariate continuous ranked probability score or its multivariate version, the energy score; see, e.g., Gneiting and Raftery (2007) and Gneiting et al. (2008) for further discussion and Warne (2022) for a recent real-time application.

h-step-ahead density forecasts is given by

$$S^{(i)}_{T:T_h,h} = \sum_{t=T}^{T_h} \log\left(p^{(i)}_{t+h|t}\right), \quad i = 1, \ldots, M. \tag{1}$$

The larger the log predictive score is, the better a model can predict the vector of variables x at the h-step-ahead forecast horizon.

A Kalman-filter-based approach to the estimation of the log predictive likelihood in linear state-space models was suggested in a recent paper by Warne et al. (2017). The basic idea is to calculate the predictive likelihood of a model conditional on a draw from the posterior distribution of the parameters and then average these likelihoods over all or a suitable subsample of the posterior draws. This simple Monte Carlo integration approach was also utilized in McAdam and Warne (2019), where we compare real-time density forecasts for the euro area based on three estimated DSGE models.

Let $w_{i,h,t}$ be the weight on model i for h-step-ahead forecasts in period t, satisfying $w_{i,h,t} \geq 0$ and $\sum_{i=1}^{M} w_{i,h,t} = 1$. The log predictive score of a generic density forecast combination using these weights is given by

$$S_{T:T_h,h} = \sum_{t=T}^{T_h} \log\left(\sum_{i=1}^{M} w_{i,h,t} p^{(i)}_{t+h|t}\right). \tag{2}$$

In the Online Appendix, Section E, we discuss several approaches to combining the density forecasts from individual models: static optimal and dynamic prediction pools (SOP and DP), Bayesian and dynamic model averaging (BMA and DMA), and log score and average log score weighting (LS and ALS). It is interesting to note that the LS weights, suggested by Jore et al. (2010), are identical to the BMA weighting approach we employ under equal initial model weights, as we shall assume in the empirical analysis. Furthermore, the ALS weights are derived from the Kullback-Leibler information criterion (KLIC) weighting scheme suggested by Mitchell and Hall (2005). It may furthermore be noted that these combination schemes cover the three broad combination methodologies discussed by Aastveit et al. (2019): frequentist based optimized combination weights (SOP); Bayesian model averaging weights (BMA, DMA and ALS); and flexible Bayesian forecast combination structures (DP). In addition to these five combination schemes, the empirical part of the paper includes equal weights (EW).

Notice that the predictive likelihood, $p(x^{(a)}_{t+h}|\mathcal{I}^{(i)}_t, A_i)$, does not include the parameters of the model as these have already been integrated out by accounting for the posterior distribution. Waggoner

and Zha (2012) allow the combination weights to follow a hidden Markov process and emphasize the joint estimation of the weights and the parameters of all models. In the empirical exercise we extend the predictive likelihoods estimates of three DSGE models from McAdam and Warne (2019) by employing vintages covering 2015–2019, while the predictive likelihoods of the reduced form models are also estimated separately. This decision is based not only on computational constraints but also on the fact that different models are rarely developed by the same team at policy institutions, making cross model estimation an unusual or unrealistic feature in practice; see also the discussion in Del Negro et al. (2016, pages 392–393).

Several other density forecast combination methods have recently been introduced to the literature, such as the dynamic Bayesian predictive synthesis in McAlinn and West (2019); the so called generalized density forecast combinations of Kapetanios et al. (2015); and the state-space approach of Billio et al. (2013); see also Aastveit et al. (2019) for additional approaches. These techniques are very general and we have opted to omit them from the current study since our objectives are mainly to learn: (i) if finite mixture combination methods can provide superior density forecasts to those from the three DSGE models in McAdam and Warne (2019) when complemented with commonly applied reduced form models; (ii) if successful methods share specific properties; and (iii) how the weights on the models develop over time and for different forecast horizons, before, during and after the Great Recession. This does not rule out that other combination schemes or models are superior to those we investigate, but we leave this issue open for future research.

2.2. Observation and Information Lags

From a recursive perspective, the weights in (2) can only be estimated based on the predictive likelihoods that have been observed at the time. The standard assumption for discrete time data is that variables are observed in the same period that they measure, i.e., x_t is both realized and observed in period t. From a real-time perspective, however, a first release or first estimate of x_t is often not available in period t but is published at a later date. Let $x_t^{(\tau)}$ denote the value of x_t taken from vintage τ, where $t \leq \tau$ and where the inequality is often strict.

Moreover, most macroeconomic variables are subject to revisions, due to more accurate information appearing with some delay and/or due to changes in measurement methodology. When comparing or evaluating forecasts, a decision must be made regarding which vintage to use for the actuals; see, e.g., Croushore and Stark (2001) and Croushore (2011). In principle, any vintage can

be used for the actual value and common choices in the real-time literature are the first release, the annual revision and the latest vintage. Although the latest vintage may reflect actuals that for many periods have not been subject to large revisions, it may suffer from possible methodological changes to the measurements that were not known in real time.[3] Similarly, first release data is typically subject to larger revisions in comparison to, for instance, annual revisions data.

The choice of actuals is important since it represents the "true value" of the forecasted variables and therefore affects the outcome of the comparison exercise. To distinguish the data used for comparing forecasts from the data used for computing model weights for combination methods, the time difference between the date of the variable and the date of the vintage the actual value is taken from is henceforth called the *observation lag* and in the empirical exercise we use annual revisions data. This means that $x_t^{(a)}$ is taken from vintage $t+4$ such that $x_t^{(a)} = x_t^{(t+4)}$, with the consequence that the observation lag $k = 4$. Similarly, if T^* denotes the latest vintage and the actuals are only taken from this vintage, then $x_t^{(a)} = x_t^{(T^*)}$ for all t and the observation lag is $T^* - t$, i.e. decreasing linearly in t. To simplify notation, we do not include a time index for the vintage date of the actuals.

At the same time, the vintages $\tau = t+1, t+2, t+3$ may include data on the forecasted variables and these measured values may be useful when computing the weights at τ. To account for this we also define the term *information lag*, denoted by l. The minimum information lag is determined by the time delay before the first publication of x_t, while the maximum may be set equal to the observation lag. The information lag is given by the difference between the vintage period, τ, and the last data point used at τ for computing combination weights, $x_{\tau-l}^{(\tau)}$. Note that the information and observation lags are both zero if one assumes that the date of the variable is equal to the time period when it is observed, as is standard for the single database (vintage) forecast exercises.

It should be emphasized that the information lag is a distinct concept from the ragged edge of real-time data; see Wallis (1986). The latter is a property of the database and is a consequence of individual time series in a real-time vintage being measured up to different time periods. For instance, interest rates may be measured up to the vintage date, while real GDP growth lags with one quarter, and some labor market variables such as wages with two quarters. The ragged edge directly affects the data available for estimation of model parameters and the conditioning information when

[3] This is certainly true for the euro area Real Time Database (RTD), which also reflects a time-varying country composition, where seven EU member countries have been admitted since 2007.

forecasting with the models.[4] The *minimum* information lag is determined by the ragged edge since it depends on the dates for which historical values of all the forecasted variables are available. At the same time, the information lag concerns only the forecasted variables and may be selected by the user of the combination method. It is not written in stone that the minimum information lag based on the ragged edge is always the best choice. A good choice for this parameter is likely to depend on the properties of the combination method as well as the size of the revisions between the initial release and the choice of actuals. For instance, the success of a combination method that has a low (high) variability of the weights over time may be expected to have a low (high) sensitivity to the information lag. Furthermore, if the revisions to the forecasted data are small, then a lower information lag is expected to be more successful than a higher lag. While we do not investigate the optimal or best selection of the information lag from a theoretical perspective, this may be an interesting topic for future research.

To clarify the relevance of these concepts and the decision problems implied by them, consider the following example based on one-quarter-ahead density forecasts: Suppose the minimum information lag is equal to one quarter for the vector x in vintage τ, while the observation lag is equal to four quarters. This means that $x_{\tau-1}$ is represented by a measured value $x_{\tau-1}^{(\tau)}$ for vintage τ and that similarly $x_{\tau-2}, x_{\tau-3}, \ldots$ have measured values for this vintage. By having a measured value it is understood that there are not any missing data for any element of x. Furthermore, an observation lag of four means that actual values of $x_{\tau-4}, x_{\tau-5}, \ldots$ are available at τ and are taken from vintages $\tau, \tau-1, \ldots$. Similarly, forecast densities of the current and all previous one-quarter-ahead forecasts of x are available for the M models at time τ, where each model makes use of data from the corresponding vintage. This means that the predictive likelihood values based on the actual values $p(x_{t+1}^{(a)}|\mathcal{I}_t^{(i)}, A_i)$ can be observed for $t = T, \ldots, \tau-5$. In addition, the predictive likelihood values based on the measured values $p(x_{t+1}^{(\tau)}|\mathcal{I}_t^{(i)}, A_i)$ can be observed for $t = T, \ldots, \tau-2$.

The user of our generic combination method needs to make two decisions before computing the weights for vintage τ: (i) which information lag to use among $l = 1, 2, 3, 4$; and (ii) whether to use the predictive likelihoods based on the actual values or on the measured values for $t = T, \ldots, \tau-4$. The decisions to these two issues determine the objective function for computing the weights. Since the actual values represent the "true values" we assume in the empirical exercise that the second

[4] Table I.2 in the Online Appendix shows the ragged edge for the real-time data for all vintages used in the empirical part. That determines which information is available when forecasting for each vintage, both when using the models individually and when computing the weights for combining the model forecasts.

decision always brings the predictive likelihood values for the actuals to the weight problem. If an information lag longer than the minimum possible is selected, this means for our example that all $x_{\tau-j}^{(\tau)}$ for $j < l$ along with the predictive likelihood values based on these measured values are discarded when computing the weights at τ. For the possible choices of l the log predictive score when computing weights is

$$\tilde{S}_{T:\tau-l-1,1} = \sum_{t=T}^{\tau-5} \log \left(\sum_{i=1}^{M} w_{i,1,t}^{(\tau)} p\big(x_{t+1}^{(a)} | \mathcal{I}_t^{(i)}, A_i\big) \right) + \sum_{t=\tau-4}^{\tau-l-1} \log \left(\sum_{i=1}^{M} w_{i,1,t}^{(\tau)} p\big(x_{t+1}^{(\tau)} | \mathcal{I}_t^{(i)}, A_i\big) \right),$$

where the weights $w_{i,1,t}^{(\tau)}$, $i = 1, \ldots, M$, are non-negative and sum to unity. Notice that the first term on the right hand side involves a sum up to the vintage date (τ) minus the observation lag ($k = 4$) and the forecast horizon ($h = 1$), while the sum for the second term begins at the vintage date minus the observation lag (plus the forecast horizon minus 1) and ends at the vintage date minus the information lag (l) and the forecast horizon. Notice also that when $\tau \leq T + l$ the above log predictive score when selecting weights cannot be determined. Initial values for the weights are then needed for these vintages and one candidate is $w_{i,1,t}^{(\tau)} = 1/M$ for these dates and for all i.

The notions of an observation lag and an information lag are indirectly recognized in the literature, but to our knowledge have not been formally conceptualized. The former is indirectly defined as it follows from the choice of actuals. Concerning the information lag, an example is given by Jore et al. (2010, p. 622) who note that "the publication delay in the production of real-time data ensures that macroeconomic variables dated through to quarter τ are not available until (vintage) $\tau+1$" and they mention that as a consequence the one-step-ahead forecasts are actually nowcasts. Furthermore, they include a one-period lag, the minimum information lag, relative to the forecast horizon when constructing their recursive weights. Note that also the information lag need not be constant. For instance, the ragged edge is not always identical for all vintages of a real-time database with the consequence that the minimum information lag can vary with the shape of the edge.

In the empirical exercise we initially make the mechanical assumption that the information lag is equal to the observation lag, but also examine the more realistic case when the information lag is shorter than the observation lag. For the variables we forecast in the empirical analysis, the shortest possible information lag for quarterly data is one quarter for most vintages and two quarters for the remaining ones. Finally, while we apply the two lag concepts to density forecast combinations in this paper, they are also relevant for point forecast combination methods with real-time data.

3. The DSGE and the VAR Models

3.1. The DSGE Models

We use three DSGE models, where the first is that of Smets and Wouters (2007), as adapted to the euro area (labelled SW). The model contains a continuum of utility-maximizing households and profit-maximizing intermediate good firms who, respectively, supply labor and intermediate goods in monopolistic competition and set wages and prices. Final good producers use these intermediate goods and operate under perfect competition.

The model incorporates several real and nominal rigidities, such as habit formation, investment adjustment costs, variable capital utilization and Calvo staggering in prices and wages. The monetary authority follows a Taylor-type rule when setting the nominal interest rate. There are seven stochastic processes: a TFP shock; a price and a wage markup shock; a risk premium (preference) shock; an exogenous spending shock; an investment-specific technology shock; and a monetary policy shock. The observed variables are: real GDP, real private consumption, real investment, employment, real wages, the GDP deflator (all transformed as 100 times the first difference of the natural logarithm), and the short-term nominal interest rate in percent.

The second model (SWFF) adds the financial accelerator mechanism of Bernanke et al. (1999) to the SW model and augments the list of observables to include a measure of the external finance premium in percent; see McAdam and Warne (2019) for details.[5] The final model (SWU) instead allows for an extensive labor margin, following Galí et al. (2012), and accordingly adds the unemployment rate in percent to the set of observables.

3.2. The VAR Models

Two BVAR models with homoskedastic innovations are studied in this paper and make use of the priors discussed in Giannone et al. (2015, 2019). The observed variables of these models are identical and, hence, they only differ in terms of their assumed priors. The first model is based on Giannone et al. (2015) with a Minnesota prior combined with the standard sum-of-coefficients prior by Doan et al. (1984), and the dummy-initial-observation prior by Sims (1993). As pointed out by Sims and Zha (1998), the latter part of the prior was designed to neutralize the bias against cointegration

[5] A variant of the SWFF, which also includes an observable for long-term inflation expectations is used by, e.g., Del Negro and Schorfheide (2013), Del Negro et al. (2015) and more recently by Cai et al. (2021).

due to the sum-of-coefficients prior, while still treating the issue of overfitting of the deterministic component; see also Sims (2000). This parameterization is henceforth called the SoC prior.

The second parameterization of the prior is based on the prior for the long run (PLR) suggested by Giannone et al. (2019). The PLR provides an alternative to the SoC prior for formulating the disbelief in an excessive explanatory power of the deterministic component of the model. Specifically, the PLR focuses on long-run relations, stationary as well as non-stationary, where economic theory can play an important role for eliciting the priors. The PLR does not impose the long-run relations but instead allows for shrinkage of the VAR parameters towards them. The details on the prior and posterior distributions as well as the estimation of the predictive likelihood are provided in the Online Appendix, Section A; (see also McAdam and Warne, 2020).

It should be stressed that the estimation approach we employ for the two BVARs with homoskedastic innovation is based on complete datasets, i.e., when there are no missing observations of the observable variables. As emphasized in the previous section, the real-time data vintages we use in the paper have a ragged edge, with some variables being missing for the vintage date as well as for the quarter prior to the vintage date. To incorporate such datasets makes direct sampling of the VAR parameters impossible and further complicates the posterior analysis as an analytical expression of the marginal likelihood conditional on the hyperparameters is not available, with the effect that all these parameters need to be estimated simultaneously. The computational costs of dealing with the ragged edge can therefore be very high and for this reason a second best approach is considered, where the dataset is trimmed during the parameter estimation step. For the forecasting step, the ragged edge is taken into account by applying a Kalman filter to the backcast, nowcast and forecast periods. The technical details on the forecasting step are presented in Section B of the Online Appendix.

A third BVAR model is also studied which allows for heteroskedastic innovations through a standard stochastic volatility setup. The modelling approach we use relies greatly on Cogley and Sargent (2005) as it has been implemented in the BEAR Toolbox; see Dieppe et al. (2016). The technical details on estimation with stochastic volatility using the above setup are given in Dieppe et al. (2018), concerning their so-called *standard model* in Section 5.2. In Section C of the Online

Appendix, we discuss the details on the prior we have used for this model as well as forecasting considerations. This BVAR model is henceforth referred to as the SV model.[6]

4. Comparing the BVAR Models to the DSGE Models

The log predictive scores of the *joint* density forecast of real GDP growth and GDP deflator inflation of the models based on the full sample of vintages are provided in Table 1.[7] For each horizon (h) and pair of rows, the log score values shown in the top row are in deviation from the largest value, while the row below gives the largest value in the column of the corresponding model. It should be kept in mind that the full sample log scores are calculated from $T_h = 76 - h$ forecast sample quarters per model for $h \geq 0$, while the backcasts involve only 3 such periods.

The full sample results show that the SoC or the SV model has the highest log score for all horizons. Moreover, all three BVAR models have a higher log score for all horizons than the best performing DSGE model. The SWU model has the highest log score among the DSGE models for all horizons. The SWFF model generally performs markedly worse than the other models, although it tends to perform comparatively better for the longer horizons.[8] Formal test results on the equality of the log predictive scores (of the different pairs of models) using the weighted likelihood ratio test, advocated by Amisano and Giacomini (2007), are shown in the Online Appendix, Figure I.5.

Recursive estimates of the average log scores for the joint density forecasts of real GDP growth and inflation for selected horizons are shown in Figure 1. Each chart displays the results for the six models for a given horizon with the SW model being represented by a solid red line, the SWFF model with a dark blue dash-dotted line, the SWU model with a green dashed line, the SoC model with a rose pink dash-dotted, the PLR with a light blue dashed line, and the SV model with a dark yellow solid line. In addition, two dark red lines called "bounds" are shown in each chart and we shall define them in Section 5.1.

---

[6] Some details on the estimation procedure of the DSGE and BVAR models is located in the Online Appendix, Section G.

[7] In a recent paper, Krüger et al. (2021) show that the log score is consistent when using MCMC parameter draws under stronger conditions than for the continuous ranked probability score and the Dawid-Sebastiani score. The latter is constructed from a Gaussian density with mean and covariance given by the predictive mean and covariance. For the six models in the current paper, the log score is estimated from a Gaussian density conditional on the parameters and averaged over the posterior parameter draws. Using Kolmogorov-Smirnov tests, it is shown by Warne (2022) that the log score and the Dawid-Sebastiani score are not statistically different for the SWU model using the same sample as in the current paper. The other two DSGE models obtain similar results compared with the Dawid-Sebastiani score in McAdam and Warne (2019) for the sample until 2014Q4. Hence, the theoretical objections one can raise when using the log score do not appear to be important for the three DSGE models and the sample investigated here.

[8] The full sample log predictive scores of the *marginal* density forecasts of the two variables are shown in the Online Appendix, Table I.4.

The horizontal axis in the panels represents the dating of the predicted variables, while the average log predictive score for a model in that period is based on all the vintages dated up to h quarters prior to the date. Concerning the DSGE models it can be seen that the paths look similar with the SWFF model path shifted down from the other two. The average log score for each DSGE model is fairly constant with a downward shift in 2008Q4. The BVAR models, on the other hand, display an upward trending behavior for the shorter horizons until 2008Q4, when a large drop occurs, before the upward trending path begins again in the aftermath of the Great Recession.

The most striking feature is the size of the drop in average log score of the BVARs compared with the DSGEs. The SoC and PLR models lose roughly twice as much in average log score as the DSGE models, while the loss for the SV model is even larger. For example, for the vintage 2008Q3 the one-quarter-ahead log predictive likelihood value for the SW and SWU models are close to -4 log units, while they are less than -11 for the BVARs. The two-quarter-ahead log predictive likelihood is around -7 for these two DSGE models and below -17 for the BVARs. To evaluate how big the losses for the BVARs are, these numbers may be compared, for the same horizons, with the log scores in Table 1, i.e. the accumulated log predictive likelihoods for all the 76 vintages in the forecast sample. In fact, the relative losses for the BVAR models are such that the model ranking changes from the BVAR models obtaining a higher average log score than the DSGE models, to the SW and SWU overtaking all three BVARs. Furthermore, from the nowcast until the eight-quarter-ahead horizon, the SV model is the best performer until the onset of the Great Recession in 2008Q4. Around 2014–15 the BVARs catch up with the SW and SWU models and thereafter overtake them.

Moving to the recursive average log scores for the real GDP growth density forecasts in Panel A of Figure 2, the pattern in connection with the onset of the Great Recession is again present. The loss in average log predictive score for the BVAR models is around one log unit, while it is a little less than half a log unit for the DSGE models. As a consequence and similar to the evidence from the joint log scores, the BVAR models temporarily lose their top rankings to the SW and SWU models.

Turning to the recursive average log scores for the inflation density forecasts in Panel B of Figure 2, the DSGE models sometimes perform better than or no worse than the BVAR models, especially for the one-quarter to four-quarter-ahead horizons. It is also notable that there is little or no effect on the short-term density forecasts from the drop in inflation during the first half of 2009, while the longer-term forecasts display a visible, albeit modest, drop in average log score for all models except the SWFF model, which includes the BGG type of financial frictions. This result is particularly

interesting since the BVAR models have access to the same data on the external finance premium, yet they are unable to utilize this information as fruitfully as the SWFF model does when forecasting inflation over 2009Q1–Q2, even two-years-ahead.

To analyze what may underlie the large drop in log score of the BVAR models relative to the DSGE models, Table 2 provides prediction errors (PEs), predictive variances (PVs) and log predictive likelihoods (LPLs) over the various horizons when the objective is to predict real GDP growth in 2008Q4 and in 2009Q1. Since the log predictive likelihood is expected to be well approximated by a Gaussian likelihood function,[9] the cause for the large drop in log score is due to prediction errors, the predictive variance or, possibly, both.

In the case of 2008Q4, the sizes of the prediction errors for the BVAR models and the DSGE models are similar in size, with all models greatly over-predicting actual quarterly real GDP growth. Moreover, there are no major differences between the prediction errors based on the vintage underlying the forecast, especially in the case of the BVARs. For example, the SoC model forecast in 2006Q4 ($h = 8$) of 2008Q4 is roughly of the same magnitude as the forecast made in 2008Q3 of 2008Q4 ($h = 1$). The only possible exception concerns the SWFF model, which has larger errors in absolute terms than the other models. Turning to the predictive variances, the estimates from the BVAR models are around three times smaller (or even more in the case of the SV model) than those from the DSGE models. Hence, the considerably smaller log predictive likelihoods of the BVAR models in 2008Q4 seems to be due to their comparatively narrow predictive densities.

Concerning real GDP growth in 2009Q1, the same explanation is supported by the estimates in Table 2. Overall, the prediction errors are larger than in 2008Q4 while the predictive variances are broadly unchanged, with the consequence that the log predictive likelihoods are much smaller for this quarter. Nevertheless, the explanation for the much larger drop in log predictive score for the BVARs than for the DSGE models is the predictive variance. The small predictive variances of the BVARs are beneficial in terms of log score prior to the Great Recession since the prediction errors are modest. However, the punishment is also severe when these models fail to predict large changes

[9] Based on the evidence presented in McAdam and Warne (2019), the predictive likelihood of each one of the three DSGE models is well approximated by a normal density based on the prediction error and the predictive variance, albeit that the approximation error is larger when the value of the log predictive likelihood is smaller. Similar results were also obtained in Warne et al. (2017) when comparing a DSGE model to DSGE-VARs and, in particular, a BVAR model based on the methodology in Bańbura et al. (2010). The posterior predictive densities of the SW, SWFF and SWU models for the real GDP growth forecasts using the 2007Q1 vintage along with the normal approximation of the predictive densities are displayed in Figure I.27 of the Online Appendix.

to the variables of interest. Still, the lower predictive variances of the BVARs is also the reason why these models recover their losses relative to the DSGE models once the Great Recession is over.[10]

5. Density Forecast Combinations

5.1. Upper And Lower Bounds for Finite Mixture Density Forecast Combinations

Forecast combinations offer an opportunity for improving upon the density forecasts of the individual models. Commonly used combination methods are typically finite mixtures where each model has a nonnegative weight and where the weights sum to unity. All methods applied in this paper belong to this category of combinations. Some combinations, such as the dynamic Bayesian predictive synthesis in McAlinn and West (2019), are compound distribution methods involving latent variables and the bounds presented below do not apply to them. Still, they are nevertheless of interest also for these methods since they provide the limits for finite mixtures.

An indicator that finite mixture combinations may be useful is that the recursive density forecasts of the individual models are not dominated by one model. The joint real GDP growth and GDP deflator inflation forecasts display time varying top ranks among the six models and similarly for real GDP growth. Concerning the inflation density forecasts, however, some horizons have a dominant model with respect to the recursive log score throughout the forecast sample; see, e.g., the eight-quarter-ahead forecasts in Panel B of Figure 2 where SV is in such a position. Nevertheless, this model does not dominate the other models in terms of log predictive likelihood for each time period, and it is therefore possible, albeit difficult, for a combination scheme to outperform the SV model.

The model weights for any finite mixture combination method are formed using information available at the time the density forecast is made. Given the models at hand, what is the best result that can be obtained by combining them? Likewise, we may ask: what is the worst result that can occur? The answers to these questions give the user an upper and a lower bound for finite mixture combinations based on the compared M models.

[10] A formal analysis of how well calibrated the model-based density forecasts are is provided in the Online Appendix, Section D, where we use the test proposed by Amisano and Geweke (2017) as well as an informal graphical analysis; see, in particular, Table I.9 and Figures I.3–I.4. The results indicate that the marginal real GDP growth density forecasts are *not* well calibrated, while the marginal inflation density forecasts, with the exception of the SWFF model, may be well calibrated.

It is straightforward to construct both *ex post* bounds when the log score is used as the scoring rule.[11] Specifically, the upper and the lower bound for each forecast horizon is obtained by collecting the maximum and the minimum of the log predictive likelihoods of the M models in each time period and adding these "optima" as the log score of the upper and the lower bound combination, respectively. That is, the upper and the lower bound of the log scores are:

$$\begin{aligned} S^{(U)}_{T:T_h,h} &= \sum_{t=T}^{T_h} \max_{i=1,\ldots,M} \log\left(p^{(i)}_{t+h|t}\right), \\ S^{(L)}_{T:T_h,h} &= \sum_{t=T}^{T_h} \min_{i=1,\ldots,M} \log\left(p^{(i)}_{t+h|t}\right). \end{aligned} \tag{3}$$

From the perspective of a forecaster combining models in real time, these bounds are, as T_h increases, close to probability zero events as they involve always picking the winner or the loser. They nevertheless form natural benchmarks when comparing density forecasts for a given set of models and forecast sample. The interval between the bounds gives the range of possible log score values that *all* finite mixture combination methods using the same models and forecast sample will take. Moreover, the difference between the upper bound and the log score of the best model is the interval available to combination methods for improving on the model forecasts. Should this interval be "too narrow", it may be prudent to consider additional forecasting models or compound methods before carrying out a combination exercise.

Returning to Figure 1, we find that the upper bound of the average log score lies quite close to the average log score of the best performing models prior to the Great Recession. Given the six models, the room for improving the density forecasts of the BVARs with combination methods up to 2008Q3 is therefore very narrow. Similarly, the lower bound of the average log score is not very far below the worst performing model (SWFF) up to this event. Once the Great Recession occurs, however, the gap between the upper bound and the best performing model increases substantially, while the gap between the SWFF model and the lower bound becomes similarly pronounced. The last two columns of Table 1 provide the upper and lower bounds for the full sample log predictive score of combination methods. Apart from the backcasts, which are based on only three data points, the best performing model has a log score approximately two-thirds-up from the lower bound for

[11] Upper and lower bounds can also be produced for other scoring functions, such as the CRPS and the energy score, using the procedure in this section, but replacing the log predictive likelihood with the relevant time period and forecast horizon score.

$h = 0, \ldots, 4$ and gradually drops thereafter to slightly less than 60 percent at $h = 8$. Hence, the room for combinations to improve upon the models' density forecasts is considerable.

Concerning the marginal density forecasts in Figure 2 we likewise find that the possibilities for combination methods to improve upon the forecasts of the best models are noteworthy for the real GDP growth forecasts after the onset of the Great Recession, while the distances between the upper bound and the best performing model for the inflation density forecasts are smaller. In addition, the inflation forecast show few changes in first rank among the models. These two aspects suggest that for inflation it will be very difficult for the combinations to beat the best performing models.

5.2. Comparing the Models to the Combination Methods

As mentioned in Section 2, six combination methods are applied for the DSGE and BVAR models and, as the default value, we set the information lag equal to the observation lag of four quarters ($k = 4$) for the SOP, the DP, the BMA, the DMA and the ALS combinations; see the Online Appendix, Section E for details. Since data releases of the predicted variables are available prior to the annual revision data release, albeit with at least one lag, we shall also examine the case when the information lag is exactly one quarter in Section 5.4.

Concerning the dynamic prediction pool, the δ^* parameter governing the effective sample size during the selection step is given by 0.90 in the Bayesian bootstrap filter. This means that an effective sample size below 90 percent of the number of particles (N) results in resampling during the selection step of the filter. The size of the latter parameter is 10,000 particles, while the grid for the ρ parameter, reflecting persistence of the dynamic pool weights, is given by $\rho \in \{0.01, 0.02, \ldots, 0.99\}$.[12] The initial values of the weights are, by assumption, equal to 1/6 for large N and these weights are used until the first time period when historical predictive likelihood values are available. With an information lag of four quarters, this occurs in period $h + 4$ of the forecast sample.

The other combination methods that allow for time-varying weights are initialized by setting them to 1/6 for each model; we analyze in some detail the importance of the initial values of the resulting log scores in the Online Appendix, Section H. Furthermore, we follow the approach in Amisano and Geweke (2017) and estimate the DMA forgetting factor, φ, and have opted to set its grid to $\varphi \in \{0.01, 0.02, \ldots, 0.99\}$.

[12] Del Negro et al. (2016) use 5,000 particles in their study with $\delta^* = 2/3$ and 10,000 posterior draws of ρ, via the random-walk Metropolis algorithm. We have checked the dynamic prediction pool results for alternative values of δ^*, namely, 0.8 and 2/3. This did not have a notable impact on the resulting log predictive scores.

The full forecast sample log predictive scores of the six models, EW, SOP, BMA, DMA and ALS combination methods are displayed in Figure 3 in deviation from the log score of the DP. The top left panel displays the results for the joint real GDP growth and inflation density forecasts, where the DSGE and BVAR models are plotted with unchanged linestyles and colors relative to the earlier graphs. The combination methods are given by the grey dashed line for EW, black dotted line for the SOP, grey solid line for BMA, black dash-dotted line for DMA and light grey solid line for ALS while the zero line represents the DP. It can be seen from the chart that all combination methods and models obtain a lower log score than the DP for all horizons, with the ALS and EW slightly behind. The differences between these three combination methods are however small.[13] The SoC and SV models also come close to the DP for the longer horizons.

Turning to real GDP growth in the top right panel, the picture is broadly similar, with the DP, ALS and EW combinations at the forefront. The BVAR models are also competitive, with the SV model doing well for the short horizons, especially the nowcasts where it comes out ahead of the combinations, and the PLR model for the medium and longer horizons. As in the joint density forecast case, the DMA approach obtains higher log scores than BMA. The SOP ranks above BMA and DMA for the nowcasts, in between them in the medium term and below them for the longer horizons.

Moving to the inflation density forecasts in the bottom left panel, an individual model is generally ranked first. For the shorter horizons, the SWU or the SoC ranks first while the SV takes this rank for the longer horizons. Among the combination methods, the SOP ranks first with BMA in second place and DMA often in third. Overall, the combination methods obtain log predictive scores within a range of 8 to -2 log units relative to the DP and for all horizons the differences are significant for the SOP, BMA and DMA methods.

To examine the behavior of the combination methods in more detail, the recursively estimated average log predictive scores of the joint density forecasts of the models and combination methods are plotted in Figure 4. To highlight the differences, the results are again shown in deviation from the recursive estimates of the average log predictive scores for the DP. Concerning ALS and EW, the deviations from zero are small, while for SOP, BMA and DMA the differences from zero follow the general pattern of the BVAR models. In view of the discussion on the upper bound in Section 5.1,

[13] Formal tests of the equality of the log scores are located in Figure I.6 of the Online Appendix. From these it can be seen, for instance, that the differences between the joint density forecasts of the DP and EW weight methods are significantly different, except for the inner forecast horizons.

all combination methods have lower log scores than the SV model up to the onset of the Great Recession and with the (temporary) downfall of the SV model these methods have an opportunity to improve the density forecasts over the individual models.[14]

5.3. Model Weights

The empirical evidence presented so far gives fairly convincing support for the usefulness of combination methods in a real-time density forecast comparison exercise. It is therefore of interest to learn how the weights on the six models develop over time and as an example we first examine the DP weights for the joint density forecasts in Figure 5.[15] In addition, summary statistics of the estimated weights for all combinations with non-fixed weights are shown in Table 3 for selected horizons.

Turning first to the estimated paths of the DP weights, recall that an information lag of four quarters is assumed. Each plot in Figure 5 therefore starts at the (large sample) initial value for $h+4$ quarters before the weights can vary. The horizontal axis of the panels represents the dating of the predicted variables, such that 2008Q4 concerns the weights used for the density forecast of real GDP growth and inflation in 2008Q4. Notice that the weights of the SV model increases when the first data on predictive likelihood values are assumed to be available, while the remaining weights either move up or down marginally. The weight on the SV model thereafter trends upward until it reaches a peak. The SV model weight then drops while the weights on the other models increase, especially the SW and SWU models. This pattern can be observed across the horizons and the largest weight on the SV model is recorded for the nowcast, which also records the largest fall.

To pinpoint where, for instance, information about 2008Q4 affects the four-quarter-ahead density forecast weights, eight periods must be added, i.e., the weight estimates for the density forecast of 2010Q4. Based on the weight paths in Figure 5, the impact of the Great Recession on the model weights is notable. However, the changes in the weights are not dramatic with all models obtaining fairly large weights throughout the forecast sample. For example, at the four-quarter horizon the combined weight of the DSGE models is around 27 percent at the end of the forecast sample, with the SWU having the largest weight and the SWFF the smallest. The slowly changing model weights

[14] The recursive estimates of the average log predictive scores of the marginal density forecasts of real GDP growth and inflation and relative to the estimates from the dynamic prediction pool are shown in Figure I.7 of the Online Appendix.

[15] The weights for the two marginal cases of real GDP growth and inflation with the dynamic pool as well as all the estimated weights based on the other methods with time-varying weights (SOP, BMA, DMA and ALS) are located in the Online Appendix, Figures I.8–I.22.

of the dynamic pool is directly related to the estimated high persistence of the underlying process with ρ being close or equal to 0.99 for most vintages.

Table 3 provides summary statistics of the DP weights for selected horizons, as well as of the other methods that support time-varying weights. It is striking how much lower the sample standard deviations of the DP weights are compared with, in particular, the SOP weights. Furthermore, the range of values (maximum minus minimum) is narrow for the DP, while the SOP frequently has the full range of possible weight values for the SV model. BMA and DMA also have large standard deviations compared with the DP and much wider ranges. Based on the summary statistics, the behavior of BMA and DMA in terms of their weights is more alike compared to the DP or the SOP, especially at the longer horizons. As might be expected, this is mainly due to the posterior estimates of the forgetting factor, φ, being close to unity for these cases. Finally, the ALS weights are often less volatile than the DP weights and their ranges tend to be narrower in several cases.[16]

To summarize, the weights of the most successful density forecast combination method over the full forecast sample, the DP and ALS, vary moderately over time, less as the forecast horizon increases, and gives substantial weight to all models. By construction, the EW method shares these properties which may explain its relative success over combination methods whose weights cover a wide range of values.[17]

5.4. The Information Lag

The combination methods depend on specific assumptions that may affect the outcome of the empirical exercises above. The information lag $l = 4$ is mechanical as it follows exactly the observation lag and it neglects the fact that the minimum information lag for real GDP growth and inflation for the RTD is often only one quarter. To simplify computational issues, we first consider the case of $l = 1$ with the measured values given by the actuals rather than taken from the corresponding vintage. Consequently, the underlying log predictive likelihoods of the DSGE and BVAR models are not re-estimated for the three additional time periods per vintage and horizon, but instead the timing of the available information is shifted backwards.[18] Specifically, the recursively obtained weights

[16] The paths for the ALS weights on the joint density forecasts of the models are different from those of the DP; see Figure I.8 of the Online Appendix.

[17] It is noteworthy that DMA approximates the equal (fixed) weights method when the forgetting factor is low. However, despite the fact that very low values of φ are allowed for when estimating this factor, the posterior mode estimates are always in the upper most part of the considered grid.

[18] In addition, all combination methods are based on having equal initial values of the model weights in one form or another. To save space, an analysis of the impact of this assumption is available in the Online Appendix, Section H.

on the models for the ALS, the BMA, the DMA and SOP are not affected by the information lag other than by shifting the weights back in time three time periods and by computing three new sets of weights at the end of the forecast sample.[19] In the case of the DP, the effect of the shorter information lag is also a simple time shift of the weights, provided that the number of particles of the underlying Bayesian bootstrap filter, N, is large enough.

Second, we also consider the case of $l = 1^*$ when all the log predictive likelihoods are estimated using the first release, the second quarter and the third quarter releases for the measured values. This allows for the use of the available information on the variables of interest in periods $\tau - 1$, $\tau - 2$ and $\tau - 3$ when computing the weights based on vintage τ data. Specifically, the measured values for period $\tau - 3$ of vintage τ correspond to the third quarter release, those for period $\tau - 2$ to the second quarter release, while the $\tau - 1$ measured values are the first release data. The main interest of this case is to investigate if using the available information at τ on measured values relative to the annual revisions actuals matters for the overall findings on the information lag or if the latter computationally cheaper case is sufficient.

Table 4 shows the full sample log predictive scores of the joint real GDP growth and inflation density forecasts for the combination methods with time-varying weights, along with the log scores for the EW method, the best performing model (BPM) and the upper bound. Turning first to the case when $l = 1$ is compared with $l = 4$, it is notable that mainly the short-term horizons are affected by the shortened information lag for all methods. In particular, substantially higher log scores are recorded for the nowcast of the BMA, DMA and SOP approaches, while the gains are smaller at the four-quarter-ahead horizon and thereafter. Furthermore, the DP or ALS typically obtains the largest log predictive score among the combination methods, with the exception of the nowcast where DMA has a larger value for $l = 1$.

The improvement in log predictive scores based on the shorter information lag is mainly due to being able to react earlier to the large forecast errors in real GDP growth (relative to the forecast error variance) recorded for the BVARs at the onset of the Great Recession.[20] This can be inferred by plotting the difference between the recursively estimated log predictive scores under $l = 1$ and

[19] The DP weights when $l = 4$ are shown in Figure 5, while those for the case $l = 1$ are visualized by lagging those weights three quarters. The weights for $l = 1$ of the other combination methods can likewise be obtained by lagging the weights in Figures I.8–I.22 three quarters.

[20] The impact on the log scores for the marginal density forecasts of real GDP growth and inflation from the information lag change are shown in the Online Appendix; see Table I.5 for the full sample and in Figure I.23 for the recursive log scores. The pattern recorded in Table 4 also applies to the log predictive scores for real GDP growth, while the log scores of inflation are only marginally affected by the shorter information lag.

$l = 4$; see Figure 6. In the case of DMA, the improvement for the *nowcast* is in excess of 10 log units for the full sample and the largest improvement occurs in 2009Q1 by having access to the log predictive likelihood for 2008Q4. As a consequence, the DMA method attaches a lower weights on the SV model, and a larger weight on the SW model at an earlier date. The SOP also gains in log score from the more timely information regarding real GDP growth at the onset of the Great Recession, although the impact on the log score is less pronounced than for the DMA.

With the exception of the nowcast and the DP, it is striking how little the recursive log score of the ALS and DP methods are affected by the information lag. From Table 3 we know that their weights are substantially less volatile than those of the other combination methods and, hence, shifting the weights back in time only has a moderate effect on the log score. This applies to all forecast horizons and also explains why they are robust to the choice of information lag for the vintages of 2001Q1 until 2019Q4. Furthermore, given the pool of models, the log scores of the models and the corresponding upper bounds, the full sample gains from using the best combination methods over the best models are noteworthy, especially for an information lag of 1.

Turning finally to the comparison of the $l = 1$ and $l = 1^*$ cases, where the former is an approximation of the latter, it is notable that the log scores are overall similar and the ranking of the combination methods is not much affected.[21] Interestingly, the differences in log predictive scores for the lower weight volatility methods DP and ALS are small. The model averaging methods as well as SOP, however, display some larger differences, especially BMA where the scores are notably higher when $h \geq 1$. Overall, however, the results suggest that the model weights are not so much affected by using the available real-time information on the measured values relative to the approximation in $l = 1$. Since the latter case uses future information when computing the weights, it cannot be used in practise, but is nevertheless of interest when comparing forecasts ex post from an information lag perspective as it greatly reduces the computational burden.

6. Summary And Conclusions

We examine finite mixture density forecast combinations of three DSGE and three BVARs across six methods: equal weights (EW), static optimal and dynamic prediction pools (SOP and DP), Bayesian and dynamic model averaging (BMA and DMA), and KLIC-based weights through the

[21] The recursively estimated average log predictive scores of the joint density forecasts for the models using the four releases as actual values are displayed in Figures I.28–I.31 in the Online Appendix. As can be seen from those graphs, the scores are not strongly affected by the choice of data release, suggesting that the combination methods will not be strongly affected by using $l = 1$ or $l = 1^*$.

average log scores (ALS). The models are estimated on real-time euro area data and the forecasts cover 2001–2019, focusing on the joint inflation and real GDP growth forecast.

In so doing, we argue that the literature on density forecast combinations has not formalized important real-time data considerations. Apart from fixed-weight combinations, model weights are computed using information about each model's past predictive performance. In a real-time context, model weighting emerges when outcomes are imperfectly known. This implies that the information set should be suitably lagged when computing the weights. To that end, we introduce the terms *observation lag* and *information lag*. The former denotes the time difference between the date of a variable and the vintage its actual value is taken from; the latter gives the time difference between the date of the vintage and the last data point of the predicted variables which is used to compute model weights when forecasting with this vintage. While the information lag affects the data available when computing model weights for predictions, the observation lag concerns the data used to compute predictive performance in a forecast comparison exercise. In the standard case of a single database both lags are zero, and this distinction is suppressed.

Furthermore, we introduce ex post based upper and lower bounds for the density forecasts for benchmarking the models and finite mixture combinations, and where the former also serves as a diagnostic aid to determine if it is worthwhile to pursue such combinations with the given models or if additional models or compound methods should be considered.

Regarding the weighting structure itself, for real GDP growth and joint forecasts with inflation, DP generally performs better than the other combination methods, with ALS and EW also obtaining competitive scores, and the individual models. For the joint forecasts, the BVAR models perform better than the DSGE models over the full sample, but they are also more sensitive to large forecast errors, such as over the Great Recession. For inflation, outcomes are more fluid with a DSGE or BVAR model typically obtaining the highest log score over the short-term and a BVAR model over the other horizons. The results for the joint forecasts are instead mainly driven by the real GDP growth forecasts, where ALS, DP and EW generally do better than the other combination methods and models. A common feature for the successful methods is a narrow range of the weight values covering the equal weights initialization.

Another dimension in which the ALS and DP are robust relates to the information lag. Shortening the lag from four quarters to one has small effects on their log scores, but strongly increases the log scores for some of the other methods and especially so for the DMA and SOP—although these

improvements abate after two or more quarter-ahead-forecasts. The gains are mainly due to being able to react earlier to the large growth forecast errors relative to the forecast uncertainty of the BVARs at the onset of the Great Recession. This illustrates the case where better models (SW, SWU) are *under*-utilized since the historical performance of the BVARs is maintained by the long information lag. The treatment of this lag, therefore, reveals a trade-off as to the degree to which combination schemes react quickly or slowly to the most recently available information.

The exercises can be extended in several directions. First, the setup of the DP involves an inherent equal-weights force through the innovation process. In one exercise, we introduce a parameterization which gives a low weight on SWFF over the initialization phase, i.e., until predictive likelihood values from the forecast sample can be observed; see the Online Appendix, Section H. One may also allow for model-dependent persistence for the weights as an alternative to the common persistence parameter. Second, compound methods have not been examined and may prove very useful. For instance, it would be interesting to examine if the dynamic Bayesian predictive synthesis approach of McAlinn and West (2019) can outperform the mixture methods and improve upon the upper bound in the euro area context. Finally, it might be interesting to extend our analysis to additional models, such as to dynamic factor models. Since our focus is not to find the best forecasting model or combination method but to learn about which properties are shared by successful finite mixture combination methods we leave these topics for future research.

References

Aastveit, K. A., Mitchell, J., Ravazzolo, F., and van Dijk, H. K. (2019), "The Evolution of Forecast Density Combinations in Economics," *Oxford Research Encyclopedia, Economics and Finance*, April, DOI: 10.1093/acrefore/9780190625979.013.381.

Amisano, G. and Geweke, J. (2017), "Prediction Using Several Macroeconomic Models," *The Review of Economics and Statistics*, 99(5), 912–925.

Amisano, G. and Giacomini, R. (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business & Economic Statistics*, 25(2), 177–190.

Bańbura, M., Brenna, F., Paredes, J., and Ravazzolo, F. (2021), "Combining Bayesian VARs with Survey Density Forecasts: Does It Pay Off?" ECB Working Paper Series No. 2543.

Bańbura, M., Giannone, D., and Reichlin, L. (2010), "Large Bayesian Vector Auto Regressions," *Journal of Applied Econometrics*, 25, 71–92.

Berg, T. O. and Henzel, S. R. (2015), "Point and Density Forecasts for the Euro Area using Bayesian VARs," *International Journal of Forecasting*, 31(4), 1067–1095.

Bernanke, B. S., Gertler, M., and Gilchrist, S. (1999), "The Financial Accelerator in a Quantitative Business Cycle Framework," in J. B. Taylor and M. Woodford (Editors), *Handbook of Macroeconomics*, volume 1C, 1341–1393, North Holland, Amsterdam.

Bernardo, J. M. (1979), "Expected Information as Expected Utility," *The Annals of Statistics*, 7(3), 686–690.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013), "Time-Varying Combinations of Predictive Densities using Nonlinear Filtering," *Journal of Econometrics*, 177(2), 213–232.

Cai, M., Del Negro, M., Herbst, E., Matlin, E., Sarfati, R., and Schorfheide, F. (2021), "Online Estimation of DSGE Models," *Econometrics Journal*, 24(1), C33–C58.

Chauvet, M. and Potter, S. (2013), "Forecasting Output," in G. Elliott, C. W. J. Granger, and A. Timmermann (Editors), *Handbook of Economic Forecasting*, volume 2, chapter 3, 141–194, Elsevier.

Clark, T. E. (2011), "Real-Time Density Forecasts From Bayesian Vector Autoregressions With Stochastic Volatility," *Journal of Business & Economic Statistics*, 29(3), 327–341.

Clements, M. P. (2017), "Assessing Macro Uncertainty in Real-Time When Data Are Subject To Revision," *Journal of Business & Economic Statistics*, 35, 420–433.

Cogley, T. and Sargent, T. J. (2005), "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII US," *Review of Economic Dynamics*, 8(2), 262–302.

Croushore, D. (2011), "Frontiers of Real-Time Data Analysis," *Journal of Economic Literature*, 49(1), 72–110.

Croushore, D. and Stark, T. (2001), "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics*, 105(1), 111–130.

Dawid, A. P. (1984), "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society, Series A*, 147(2), 278–292.

Del Negro, M., Giannoni, M. P., and Schorfheide, F. (2015), "Inflation in the Great Recession and New Keynesian Models," *American Economic Journal: Macroeconomics*, 7(1), 168–196.

Del Negro, M., Hasegawa, R. B., and Schorfheide, F. (2016), "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance," *Journal of Econometrics*, 192(2), 391–403.

Del Negro, M. and Schorfheide, F. (2013), "DSGE Model-Based Forecasting," in G. Elliott and A. Timmermann (Editors), *Handbook of Economic Forecasting*, volume 2, 57–140, North Holland, Amsterdam.

Diebold, F., Gunther, T. A., and Tay, A. S. (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39(4), 863–883.

Dieppe, A., Legrand, R., and van Roye, B. (2016), "The BEAR Toolbox," ECB Working Paper Series No. 1934.

Dieppe, A., Legrand, R., and van Roye, B. (2018), "The Bayesian Estimation, Analysis and Regression (BEAR) Toolbox: Technical Guide," Technical Document, BEAR Toolbox, European Central Bank.

Doan, T., Litterman, R., and Sims, C. A. (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, 3(1), 1–100.

Edge, R. M., Kiley, M. T., and Laforte, J.-P. (2010), "A Comparison of Forecast Performance Between Federal Reserve Staff Forecasts, Simple Reduced-Form Models, and a DSGE Model," *Journal of Applied Econometrics*, 25, 720–754.

Galí, J., Smets, F., and Wouters, R. (2012), "Unemployment in an Estimated New Keynesian Model," in D. Acemoglu and M. Woodford (Editors), *NBER Macroeconomics Annual 2011*, 329–360, University of Chicago Press.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015), "Prior Selection for Vector Autoregressions," *The Review of Economics and Statistics*, 97(2), 436–451.

Giannone, D., Lenza, M., and Primiceri, G. E. (2019), "Priors for the Long Run," *Journal of the American Statistical Association*, 114(526), 565–580.

Gneiting, T. and Katzfuss, M. (2014), "Probabilistic Forecasting," *The Annual Review of Statistics and Its Applications*, 1, 125–151.

Gneiting, T. and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102(477), 359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008), "Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Predictions of Surface Winds," *TEST*, 17(2), 211–235.

Groen, J. J. J., Paap, R., and Ravazzolo, F. (2013), "Real-Time Inflation Forecasting in a Changing World," *Journal of Business & Economic Statistics*, 31(1), 29–44.

Hall, S. G. and Mitchell, J. (2007), "Combining Density Forecasts," *International Journal of Forecasting*, 23(1), 1–13.

Jarociński, M. and Lenza, M. (2018), "An Inflation-Predicting Measure of the Output Gap in the Euro Area," *Journal of Money, Credit and Banking*, 50(6), 1189–1224.

Jore, A. S., Mitchell, J., and Vahey, S. P. (2010), "Combining Forecast Densities from VARs with Uncertain Instabilities," *Journal of Applied Econometrics*, 25(4), 621–634.

Kapetanios, G., Mitchell, J., Price, S., and Fawcett, N. (2015), "Generalised Density Forecast Combinations," *Journal of Econometrics*, 188(1), 150–165.

Krüger, F., Lerch, S., Thorarinsdottir, T., and Gneiting, T. (2021), "Predictive Inference Based on Markov Chain Monte Carlo Output," *International Statistical Review*, 89(2), 274–301.

Mazzi, G. M., Mitchell, J., and Montana, G. (2014), "Density Nowcasts and Model Combination: Nowcasting Euro-Area GDP Growth over the 2008–09 Recession," *Oxford Bulletin of Economics and Statistics*, 76(2), 233–256.

McAdam, P. and Warne, A. (2019), "Euro Area Real-Time Density Forecasting with Financial or Labor Market Frictions," *International Journal of Forecasting*, 35(2), 580–600.

McAdam, P. and Warne, A. (2020), "Density Forecast Combinations: The Real-Time Dimension," ECB Working Paper Series No. 2378.

McAlinn, K. and West, M. (2019), "Dynamic Bayesian Predictive Synthesis in Time Series Forecasting," *Journal of Econometrics*, 210(1), 155–169.

Mitchell, J. and Hall, S. G. (2005), "Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation," *Oxford Bulletin of Economics and Statistics*, 67(S1), 995–1033.

Rossi, B. and Sekhposyan, T. (2014), "Evaluating Predictive Densities of US Output Growth and Inflation in a Large Macroeconomic Data Set," *International Journal of Forecasting*, 30(3), 662–682.

Sims, C. A. (1993), "A Nine-Variable Probabilistic Macroeconomic Forecasting Model," in J. H. Stock and M. W. Watson (Editors), *Business Cycles, Indicators and Forecasting*, 179–212, University of Chicago Press, Chicago.

Sims, C. A. (2000), "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples," *Journal of Econometrics*, 95(2), 443–462.

Sims, C. A. and Zha, T. (1998), "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, 39(4), 949–968.

Smets, F., Warne, A., and Wouters, R. (2014), "Professional Forecasters and Real-Time Forecasting with a DSGE Model," *International Journal of Forecasting*, 30(4), 981–995.

Smets, F. and Wouters, R. (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97(3), 586–606.

Timmermann, A. (2006), "Forecast Combinations," in G. Elliott, C. W. J. Granger, and A. Timmermann (Editors), *Handbook of Economic Forecasting*, volume 1, chapter 4, 135–196, Elsevier, Amsterdam.

Waggoner, D. F. and Zha, T. (2012), "Confronting Model Misspecification in Macroeconomics," *Journal of Econometrics*, 171(2), 167–184.

Wallis, K. F. (1986), "Forecasting with an Econometric Model: The Ragged Edge Problem," *Journal of Forecasting*, 5(1), 1–13.

Warne, A. (2022), "DSGE Model Forecasting: Rational Expectations vs. Adaptive Learning," Manuscript, European Central Bank.

Warne, A., Coenen, G., and Christoffel, K. (2017), "Marginalized Predictive Likelihood Comparisons of Linear Gaussian State-Space Models with Applications to DSGE, DSGE-VAR and VAR Models," *Journal of Applied Econometrics*, 32(1), 103–119.

TABLE 1: Log predictive scores for the joint density forecasts of real GDP growth and GDP deflator inflation of the six models over the vintages 2001Q1–2019Q4.

	DSGE			BVAR			Bounds	
h	SW	SWFF	SWU	SoC	PLR	SV	Upper	Lower
−1	−5.14	−6.31	−4.94	−1.45	−1.24	0.00		
						−4.06	−4.06	−10.70
0	−18.90	−61.21	−13.43	−7.66	−8.99	0.00		
						−28.82	7.15	−125.46
1	−14.90	−78.10	−8.68	0.00	−0.36	−0.74		
				−47.68			−10.54	−150.39
2	−15.23	−77.13	−8.32	0.00	−2.34	−4.44		
				−56.89			−18.38	−159.41
3	−20.13	−72.42	−11.32	0.00	−3.12	−5.33		
				−60.15			−24.93	−160.68
4	−21.08	−62.02	−12.16	0.00	−2.25	−2.36		
				−62.76			−28.31	−155.62
5	−20.59	−52.43	−11.13	0.00	−2.56	−2.55		
				−65.25			−32.21	−149.12
6	−21.91	−45.77	−12.88	0.00	−4.95	−0.66		
				−66.02			−34.63	−144.43
7	−21.94	−37.90	−14.26	−1.49	−6.06	0.00		
						−67.15	−35.69	−140.62
8	−23.50	−34.30	−17.47	−4.23	−10.20	0.00		
						−66.25	−36.10	−137.10

NOTES: The log scores are displayed in deviation from the largest value of the best performing model for each horizon. The largest log score is shown in the row below for the model which achieves this value. The backcast horizon is denoted by $h = -1$ while the nowcast horizon is given by $h = 0$. The ex-post-based upper and lower bounds for combination methods are shown in the last two columns; see equation (3).

TABLE 2: Predictions error, variance and log predictive likelihood of real GDP growth in 2008Q4 and 2009Q1.

		2008Q4						2009Q1					
		DSGE			BVAR			DSGE			BVAR		
h	Type	SW	SWFF	SWU	SoC	PLR	SV	SW	SWFF	SWU	SoC	PLR	SV
	PE	−1.85	−2.13	−2.03	−1.98	−1.93	−2.10	−1.81	−1.82	−2.74	−2.37	−2.32	−2.23
0	PV	0.45	0.46	0.42	0.15	0.14	0.15	0.48	0.50	0.43	0.19	0.18	0.29
	LPL	−4.30	−5.42	−5.32	−12.08	−12.09	−9.83	−3.95	−3.96	−8.92	−13.29	−13.80	−6.40
	PE	−1.92	−2.31	−1.99	−2.26	−2.27	−2.32	−2.69	−3.00	−2.80	−2.71	−2.69	−2.91
1	PV	0.54	0.55	0.59	0.20	0.18	0.12	0.54	0.55	0.59	0.20	0.19	0.18
	LPL	−3.98	−5.38	−4.00	−11.98	−12.66	−13.42	−7.07	−8.46	−7.12	−16.09	−16.79	−12.88
	PE	−2.11	−2.41	−2.08	−2.25	−2.26	−2.37	−2.68	−3.06	−2.72	−2.80	−2.79	−2.99
2	PV	0.56	0.56	0.64	0.22	0.20	0.13	0.56	0.56	0.64	0.21	0.20	0.12
	LPL	−4.54	−5.69	−4.04	−10.88	−11.78	−13.45	−6.79	−8.61	−6.33	−16.16	−17.18	−19.64
	PE	−2.23	−2.55	−2.17	−2.28	−2.17	−2.35	−2.92	−3.18	−2.93	−2.89	−2.82	−2.99
4	PV	0.59	0.60	0.68	0.24	0.21	0.15	0.58	0.57	0.67	0.24	0.21	0.14
	LPL	−4.85	−6.06	−4.16	−10.22	−10.25	−14.58	−7.74	−9.15	−6.96	−14.87	−16.03	−21.01
	PE	−2.29	−2.65	−2.24	−2.23	−2.10	−2.36	−2.98	−3.25	−2.86	−2.83	−2.70	−2.98
8	PV	0.63	0.63	0.71	0.26	0.23	0.19	0.60	0.60	0.72	0.26	0.22	0.18
	LPL	−4.82	−6.19	−4.29	−8.83	−9.24	−13.75	−7.78	−9.18	−6.30	−12.83	−14.23	−22.40

NOTES: The three types of predictive distribution estimates are: prediction error (PE), predictive variance (PV) and log predictive likelihood (LPL). The forecast horizon, h, determines which euro area RTD vintage is used to predict real GDP growth in 2008Q4 and 2009Q1, respectively. For example, $h = 4$ when predicting the outcome in 2008Q4 implies that the 2007Q4 vintage is employed. The actual values of quarterly real GDP growth in 2008Q4 and 2009Q1 are given by −1.89 and −2.53 percent, respectively.

TABLE 3: Summary statistics of the weights on the DSGE and BVAR models for combination methods for the joint density forecasts of real GDP growth and inflation over the vintages 2001Q1–2019Q4.

		DP				SOP				BMA				DMA				ALS			
h	model	mean	std	min	max	mean	std	min	max	mean	std	min	max	mean	std	min	max	mean	std	min	max
0	SW	0.14	0.05	0.08	0.28	0.12	0.11	0.00	0.32	0.33	0.41	0.00	0.96	0.16	0.27	0.00	0.95	0.17	0.02	0.10	0.21
	SWFF	0.10	0.03	0.06	0.17	0.01	0.04	0.00	0.17	0.01	0.04	0.00	0.17	0.02	0.04	0.00	0.17	0.11	0.02	0.08	0.17
	SWU	0.13	0.03	0.08	0.18	0.01	0.04	0.00	0.17	0.04	0.06	0.00	0.28	0.09	0.09	0.00	0.36	0.17	0.01	0.15	0.23
	SoC	0.19	0.03	0.14	0.29	0.12	0.14	0.00	0.49	0.03	0.04	0.00	0.20	0.14	0.12	0.00	0.39	0.17	0.02	0.12	0.20
	PLR	0.20	0.04	0.14	0.28	0.23	0.25	0.00	0.77	0.04	0.05	0.00	0.17	0.15	0.12	0.00	0.44	0.17	0.02	0.12	0.20
	SV	0.24	0.10	0.12	0.45	0.51	0.42	0.00	1.00	0.56	0.41	0.00	1.00	0.43	0.33	0.00	0.96	0.20	0.03	0.17	0.29
1	SW	0.14	0.04	0.08	0.22	0.02	0.05	0.00	0.21	0.21	0.21	0.00	0.59	0.14	0.19	0.00	0.62	0.18	0.02	0.15	0.21
	SWFF	0.08	0.03	0.05	0.17	0.01	0.04	0.00	0.17	0.01	0.04	0.00	0.17	0.02	0.04	0.00	0.17	0.10	0.02	0.07	0.17
	SWU	0.14	0.04	0.08	0.20	0.13	0.15	0.00	1.00	0.21	0.20	0.00	0.56	0.14	0.14	0.00	0.48	0.18	0.02	0.15	0.22
	SoC	0.20	0.03	0.14	0.26	0.07	0.16	0.00	0.70	0.14	0.16	0.00	0.55	0.17	0.11	0.00	0.49	0.18	0.02	0.11	0.20
	PLR	0.23	0.07	0.15	0.36	0.36	0.34	0.00	0.87	0.17	0.14	0.00	0.49	0.22	0.16	0.00	0.60	0.18	0.02	0.13	0.21
	SV	0.20	0.06	0.11	0.34	0.40	0.42	0.00	1.00	0.25	0.28	0.00	0.87	0.31	0.26	0.00	0.82	0.19	0.02	0.16	0.24
4	SW	0.14	0.02	0.08	0.17	0.02	0.05	0.00	0.17	0.11	0.13	0.00	0.38	0.11	0.13	0.00	0.37	0.17	0.02	0.14	0.20
	SWFF	0.11	0.03	0.08	0.17	0.02	0.05	0.00	0.17	0.02	0.05	0.00	0.17	0.02	0.05	0.00	0.17	0.12	0.02	0.09	0.17
	SWU	0.15	0.04	0.10	0.24	0.11	0.09	0.00	0.24	0.26	0.31	0.00	0.84	0.18	0.25	0.00	0.73	0.17	0.02	0.13	0.21
	SoC	0.17	0.01	0.14	0.20	0.02	0.05	0.00	0.17	0.14	0.22	0.00	0.82	0.11	0.08	0.00	0.28	0.18	0.01	0.14	0.20
	PLR	0.18	0.04	0.14	0.29	0.03	0.06	0.00	0.17	0.10	0.14	0.00	0.54	0.14	0.17	0.00	0.60	0.17	0.01	0.14	0.19
	SV	0.25	0.07	0.16	0.40	0.81	0.25	0.17	1.00	0.37	0.44	0.00	1.00	0.44	0.36	0.00	0.95	0.20	0.05	0.13	0.34
8	SW	0.14	0.02	0.11	0.18	0.03	0.06	0.00	0.17	0.19	0.20	0.00	0.61	0.11	0.11	0.00	0.51	0.17	0.02	0.14	0.21
	SWFF	0.14	0.02	0.11	0.17	0.03	0.06	0.00	0.17	0.04	0.06	0.00	0.17	0.09	0.12	0.00	0.55	0.14	0.02	0.11	0.17
	SWU	0.15	0.03	0.11	0.26	0.10	0.08	0.00	0.24	0.21	0.22	0.00	0.75	0.14	0.16	0.00	0.70	0.16	0.02	0.13	0.20
	SoC	0.16	0.01	0.13	0.17	0.03	0.06	0.00	0.17	0.08	0.09	0.00	0.34	0.11	0.07	0.00	0.27	0.17	0.01	0.11	0.18
	PLR	0.15	0.01	0.13	0.17	0.03	0.06	0.00	0.17	0.04	0.06	0.00	0.18	0.10	0.07	0.00	0.22	0.16	0.01	0.15	0.18
	SV	0.26	0.06	0.16	0.34	0.78	0.29	0.17	1.00	0.44	0.41	0.00	1.00	0.44	0.32	0.00	0.94	0.19	0.05	0.12	0.27

NOTES: The density combination methods are: dynamic prediction pool (DP), static optimal prediction pool (SOP), Bayesian model averaging (BMA), dynamic model averaging (DMA), and average log score (ALS).

TABLE 4: The impact of the information lag for the log predictive scores of joint real GDP growth and GDP deflator inflation density forecast comparisons over the vintages 2001Q1–2019Q4.

h	l	SOP	DP	BMA	DMA	ALS	EW	BPM	Upper
0	1	−24.78	−24.72	−30.61	−19.38	−27.55	−29.88	−28.82	7.15
	1^*	−26.65	−24.51	−30.58	−20.02	−27.67			
	4	−31.82	−27.11	−37.18	−30.13	−28.44			
1	1	−48.55	−40.63	−56.76	−41.49	−43.61	−47.12	−47.68	−10.54
	1^*	−49.32	−40.55	−54.16	−42.07	−43.61			
	4	−51.06	−41.97	−58.64	−48.93	−44.38			
2	1	−66.22	−50.46	−71.29	−55.85	−53.46	−56.16	−56.89	−18.38
	1^*	−66.81	−50.11	−67.92	−53.77	−53.59			
	4	−69.33	−51.70	−73.10	−58.22	−54.24			
4	1	−68.26	−58.48	−78.46	−64.43	−61.04	−63.22	−62.76	−28.31
	1^*	−68.22	−58.30	−76.00	−61.39	−60.98			
	4	−69.14	−59.54	−79.17	−66.54	−61.51			
8	1	−69.90	−65.48	−78.98	−66.46	−68.44	−69.17	−66.25	−36.10
	1^*	−72.24	−66.26	−75.54	−66.11	−68.46			
	4	−72.02	−65.78	−78.87	−66.57	−68.57			

NOTES: The case when $l = 1$ refers to letting the measured values of the predicted variables in time periods $\tau - 1, \tau - 2, \tau - 3$ for vintage τ when computing weights using an information lag of 1 be given by the actual values of the predicted variables, while $l = 1^*$ is based on using the measured values in those time periods from vintage τ. The equal weights and the upper bound log predictive scores are, by construction, invariant to the information lag, where the best performing model (BPM) is given by the individual DSGE/BVAR model with the largest log predictive score for each horizon (as taken from Table 1).

FIGURE 1: Recursive estimates of the average log predictive scores of the joint density forecasts of real GDP growth and GDP deflator inflation for the vintages 2001Q1–2019Q4.

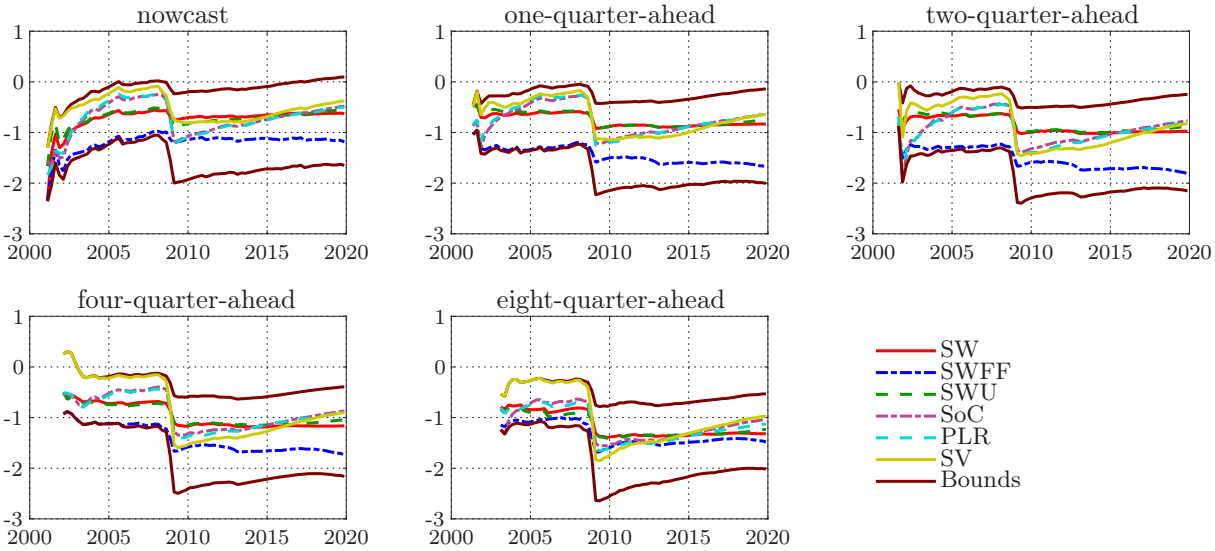


FIGURE 2: Recursive estimates of the average log predictive scores of the marginal density forecasts of real GDP growth and GDP deflator inflation for the vintages 2001Q1–2019Q4.

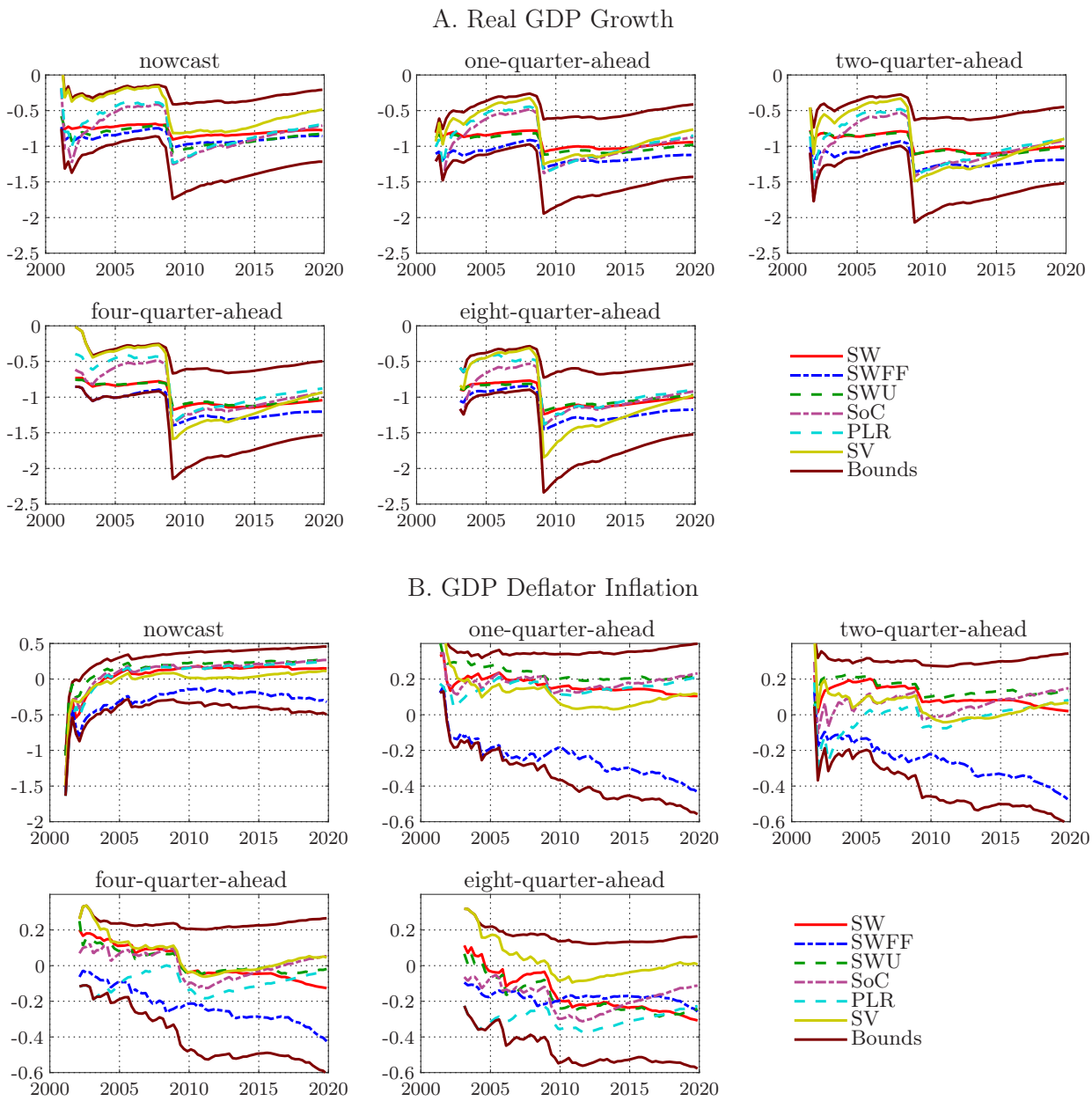
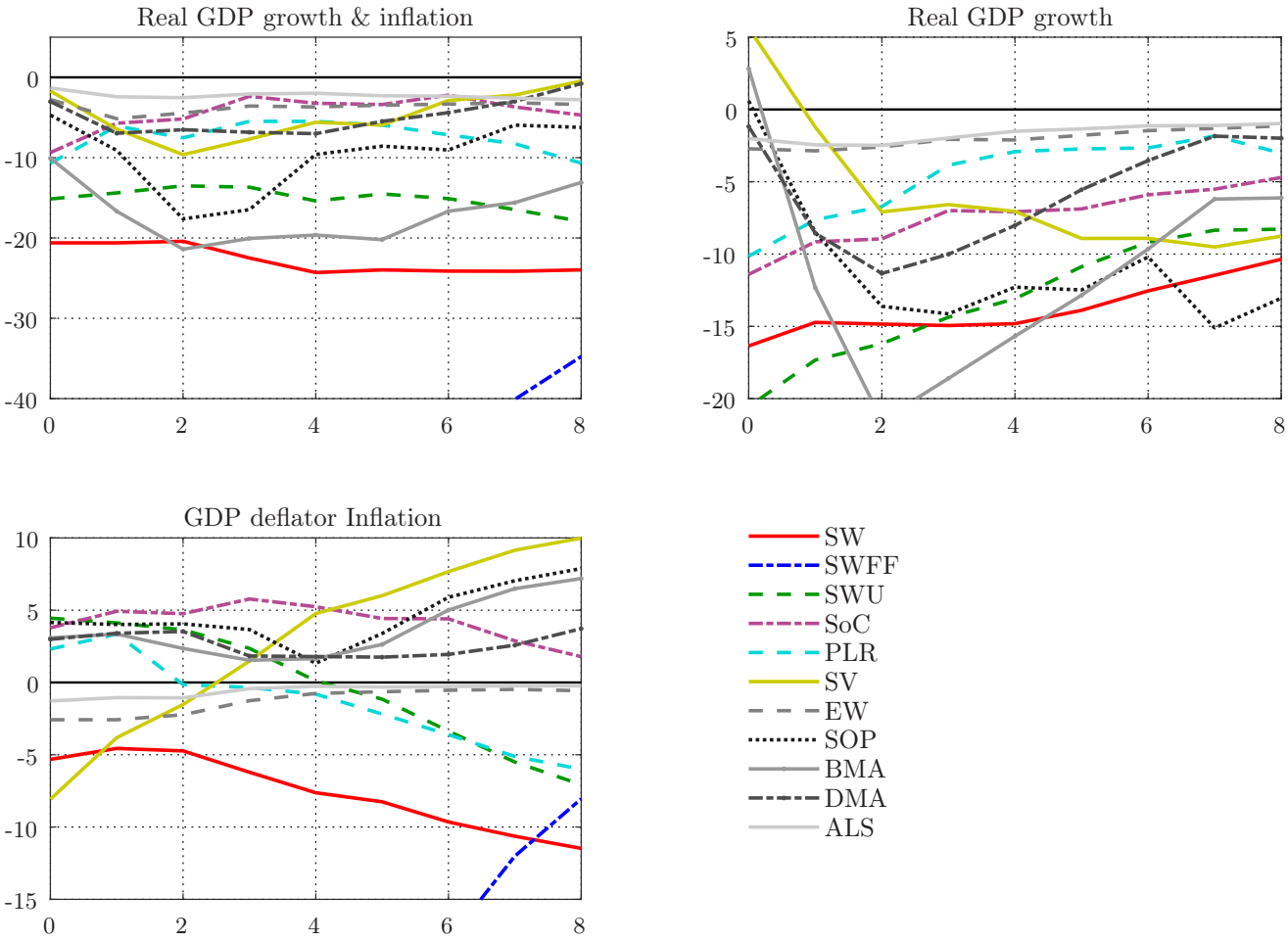


FIGURE 3: Log predictive scores for nowcasts and one-quarter-ahead to eight-quarter-ahead forecasts of DSGE models, BVAR models and combination methods in deviation from the log score of the dynamic prediction pool over the vintages 2001Q1–2019Q4.



NOTES: All full sample log predictive scores are measured in deviation from the log score of the dynamic prediction pool (DP). The other density forecast combination methods are given by equal weight (EW), static optimal prediction pool (SOP), Bayesian model averaging (BMA), dynamic model averaging (DMA) and average log score (ALS). The DP, SOP, BMA, DMA and ALS combination methods are based on an information lag of four quarters.

FIGURE 4: Recursive estimates of the average log predictive scores of the joint density forecasts of real GDP growth and GDP deflator inflation and in deviation from the recursive estimates of the average log scores of the dynamic prediction pool covering the vintages 2001Q1–2019Q4.

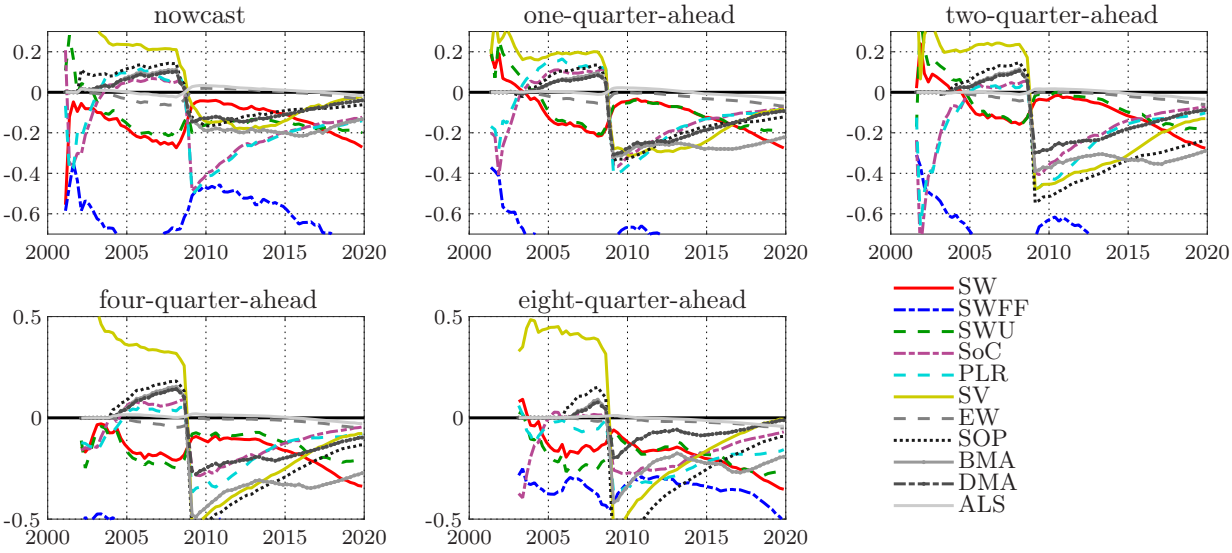


FIGURE 5: Posterior estimates of the model weights for the dynamic prediction pool of the joint density forecasts of real GDP growth and GDP deflator inflation covering the vintages 2001Q1–2019Q4.

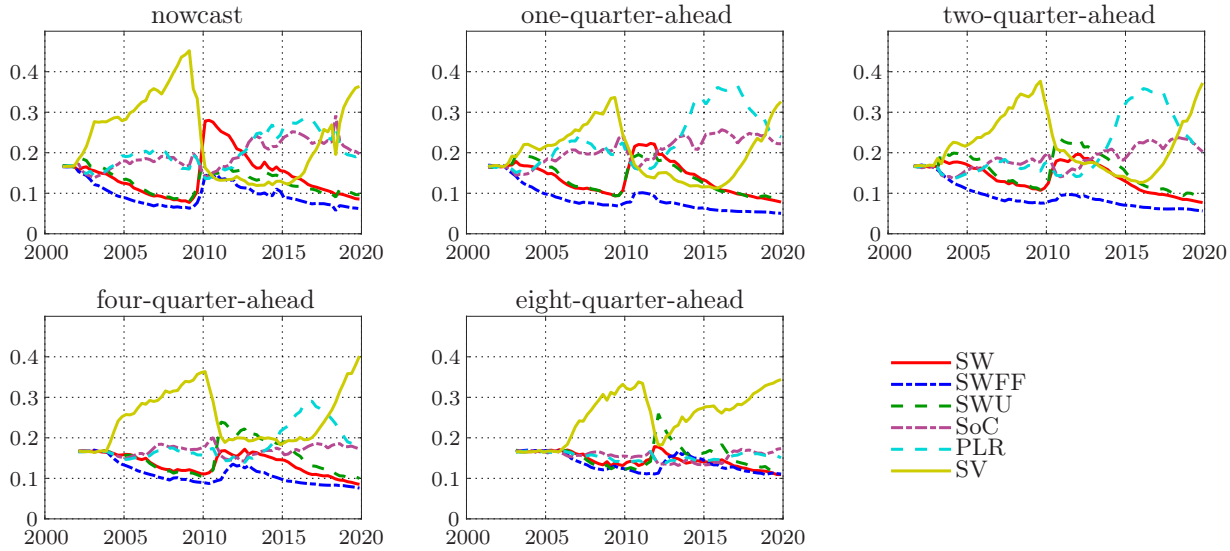


FIGURE 6: Recursive estimates of the differences of log predictive scores of joint real GDP growth and GDP deflator inflation density forecasts with information lag 1 and 4 covering the vintages 2001Q1–2019Q4.