Combining Bayesian VARs with survey density forecasts: does it pay off?*

Marta Bańbura[a], Federica Brenna[b], Joan Paredes[a] and Francesco Ravazzolo[c]

[a] European Central Bank
[b] KU Leuven and European Central Bank
[c] Free University of Bozen-Bolzano, BI Norwegian Business School and RCEA

Abstract

This paper studies how to combine real-time forecasts from a broad range of Bayesian vector autoregression (BVAR) specifications and survey forecasts by optimally exploiting their properties. To do that, it compares the forecasting performance of optimal pooling and tilting techniques, including survey forecasts for predicting euro area inflation and GDP growth at medium-term forecast horizons, using both univariate and multivariate forecasting metrics. Results show that the Survey of Professional Forecasters (SPF) provides good point forecast performance, but scores poorly in terms of densities for all variables and horizons. Accordingly, when individual models are tilted to SPF's first moments and then optimally combined, point accuracy and calibration improve, whereas they worsen when SPF's second moments are included in the tilting. We conclude that judgement incorporated in survey forecasts can considerably increase model forecasts accuracy, however, the way and the extent to which it is incorporated matters.

Keywords: Real Time, Optimal Pooling, Judgement, Entropic tilting, Survey of Professional Forecasters

JEL Codes: C11, C32, C53, E27, E37

*The authors thank Todd Clark, Michael Clements, Matthieu Darracq-Pariès, Carlos Montes-Galdón, Gergely Ákos Gánics, Marek Jarociński, Gael Martin, Elmar Mertens, James Mitchell, Pilar Poncela, Giorgio Primiceri, Ellis Tallman, Oreste Tristani, Saeed Zaman, an anonymous referee and conference and seminar participants at the 11th ECB Conference on Forecasting Techniques, the 27th International Conference Computing in Economics and Finance, the 2021 Annual Conference of the International Association for Applied Econometrics, the 7th RCEA Time Series Workshop, the 41st International Symposium on Forecasting, the 11th European Seminar on Bayesian Econometrics and the IIF's MacroFor Seminar for comments and support. All remaining errors are our own. The views expressed are those of the authors and do not necessarily reflect those of the European Central Bank (ECB).

1 Introduction

Optimally combining forecasts from multiple models in order to robustly predict future paths of macroeconomic variables is a methodology which has been advocated for some time in the economic literature (for a comprehensive review, see Bassetti et al., 2020). The reason is that it is hard to find an individual model, which can be considered the "best performing" in all possible forecasting dimensions, i.e. for any variable, at any forecast horizon, at any point in history, and for any loss function metric (be it in terms of point or density forecast). It is then quite natural to think about combinations as a way of averaging multiple measurements of the same outcome. These measurements may be the result of known econometric models, or they may come from a mixture of unobserved data, models, and judgement calls, such as the figures published in survey forecasts. The latter have been shown to improve forecast accuracy when added to model forecasts (see e.g. Krüger et al., 2017; Tallman and Zaman, 2020).

Having the above in mind, we construct a real-time forecast exercise for the euro area real GDP and HICP inflation using econometric models at one- and two-year-ahead forecast horizons. We use several types of Bayesian Vector Auto-Regression models (BVARs), which became a standard tool for forecasting and scenario analysis in the central banking community, above all for mid- and long-term forecast horizons (see e.g. Domit et al., 2016; Angelini et al., 2019; Crump et al., 2021). We choose several specifications, which differ on certain modelling choices, such as data set size and composition, data transformation, degree of time variation, prior specification, and inclusion of off-model information. In particular, we include standard BVAR models with constant coefficients with Minnesota (Sims and Zha, 1998; Bańbura et al., 2010; Carriero et al., 2019) and democratic priors (Villani, 2009; Clark, 2011; Wright, 2013); a model with time-varying parameters (Primiceri, 2005); a model with a local mean and constant coefficients (Garnier et al., 2015; Crump et al., 2016; Mertens, 2016; Del Negro et al., 2017; Bańbura and van Vlodrop, 2018); and also a univariate unobserved component model with stochastic volatility (UCSV) in the style of Stock and Watson (2007). For some models we use both a 3 and a 19 variable specification, resulting overall in 13 different BVARs. Finally, we combine those model forecasts by means of linear optimal pooling, where weights are selected in order to maximise forecast accuracy (Hall and Mitchell, 2007; Jore et al., 2010; Geweke and Amisano, 2011; Amisano and Geweke, 2017), both for each individual variable and for the aggregate of the two variables.

We evaluate the performance of individual models and their combinations over the period 2000-2019 at the one- and two-year-ahead horizons, in terms of point and density forecast accuracy by calculating Root Mean Square Forecast Errors (RMSFE), Log Predictive Scores (LPS) and Continuous Ranked Probability Scores (CRPS). In order to assess calibration, we compute the Probability Integral Transforms (PITs) and perform a test for the uniformity of the PITs distribution (Berkowitz, 2001). The latter feature is often overlooked in forecast evaluations, although it is key when accurate measures of uncertainty around the point forecasts are needed. We find that combination improves on individual models, however, it does not achieve good calibration for both variables and horizons.

We then turn to an additional source of information, namely the Survey of Professional Forecasters (SPF), whose forecasts are known for having a good point forecast performance (Ang et al.,

2007; Kenny et al., 2014). We construct a continuous distribution from the SPF histograms in order to assess their accuracy and calibration. We find, as expected, high accuracy for the SPF point forecasts, but poor performance from a density perspective, both in terms of accuracy and calibration. Similar results are found, among others, in Clements (2018) and Clements (2014) for the U.S. SPF.

As our previous analysis does not yield a clear "best" forecast when comparing a combination of BVARs against survey forecasts, we investigate two additional approaches to incorporate information from both methods in a unique forecast density, hence benefiting from more subjective SPF forecasts, containing forward-looking information, and from the more academic and model based BVAR forecasts, which are mostly based on backward looking information.

In the first approach, after simulating draws from the SPF distributions we obtain a SPF density forecast, which we incorporate into the pool of models used for the optimal combination.
The second method is entropic tilting. Namely, we tilt either the individual BVAR models or their model combination to either the first moment or both first and second moments of the SPF. Therefore, we extend the literature that applies tilting to individual models (Krüger et al., 2017; Altavilla et al., 2017; Ganics and Odendahl, 2021; Tallman and Zaman, 2020) and model combinations (Galvao et al., 2021) or just combines macroeconomic models (Amisano and Geweke, 2017). To our knowledge, we are the first to combine tilted forecasts or to include survey forecasts in macroeconomic model density combinations.

We find that incorporating survey information in model combinations improves forecast performance and calibration, especially in the tilting method, albeit only limited to the first survey moment. When the individual models or the combination are tilted to both mean and variance of the SPF, there is a general worsening of the performance. Our results are similar to Galvao et al. (2021) for U.K.'s GDP and inflation; they also found that judgement on the mean tends to improve model density forecasts at short horizons, whereas survey second moments hinder performance at short horizons. Combining individual BVARs both with and without judgement improves accuracy with respect to individual models according to the LPS metric, while in terms of CRPS the combination is worse than SPF for all variables and horizons, with the exception of the two-year-ahead GDP forecast.

The option that improves forecast performance the most, both in point and in density terms for both variables and horizons, is the combination tilted to the SPF mean "ex-ante" (before performing optimal pooling, each individual model is tilted). This higher performance is due to improved initial conditions: when models are tilted to informative moments, they perform better individually; consequently, the optimal pool will select different models compared to the non-tilted case resulting in better final scores. We conclude that improving individual models (for example by tilting each model to the SPF mean) and then combining them optimally is the best forecast strategy, highlighting the complementary role of both methods. The improved initial conditions are clearly visible when looking at scores for the non-tilted and the mean-tilted individual models relative to the optimal pool (available in the online Appendix).

For robustness, we finally extend the evaluation to a bivariate setting, where accuracy is computed using joint statistics for GDP and inflation forecasts, allowing to investigate whether models' or

surveys' performance for one of the two variables can be explained by conditioning on the other variable. Results are consistent with the univariate case, i.e. the optimal pool tilted ex-ante to the mean of SPF provides again the most accurate results. We also confirm that the lack of calibration of the SPF is not due to the survey respondents giving higher relevance to one of the two variables, but rather their densities are simply too narrow (i.e. they are over-confident) for both variables.

We conclude that judgement incorporated in survey forecasts can help model forecasts accuracy and improve calibration, however, the way and extent to which it is incorporated matters.

The rest of the paper is organised as follows. In Section 2, we present the set of BVAR models included in the combination and briefly explain the data used and combination method, as well as show the performance results of the optimal pooling combination. In Section 3, we explain how we construct the SPF forecast density from a discrete histogram and how do these forecasts perform. In Section 4, we illustrate the two options used to combine model and survey information, and present their results. Section 5 describes results for the bivariate analysis and Section 6 concludes.

2 Pool of Bayesian VARs

Bayesian VARs have become a standard tool for forecasting and scenario analysis at central banks, due to their competitive performance and relatively easy implementation. At the same time, results from such models are often sensitive to some modelling choices such as data set size and composition, data transformation, degree of time variation, prior specification, and inclusion of off-model information. To hedge against such model uncertainty, we consider several of the most common model variants. In particular, we choose standard BVAR models with constant coefficients with Minnesota (Sims and Zha, 1998; Bańbura et al., 2010; Carriero et al., 2019) and democratic priors (Villani, 2009; Clark, 2011; Wright, 2013); a model with time-varying parameters (Primiceri, 2005); a model with a local mean and constant coefficients (Garnier et al., 2015; Crump et al., 2016; Mertens, 2016; Del Negro et al., 2017; Bańbura and van Vlodrop, 2018); and also a univariate unobserved component model with stochastic volatility (UCSV) in the style of Stock and Watson (2007). The survey local mean model and the specification with democratic priors allow including off-model information (from long-term survey forecasts) to pin down the low frequency evolution of the modelled variables (the trends). We consider "small" (three variables) and "medium" (19 variables) data set sizes for the euro area as a whole and a bottom-up approach whereby the euro area forecast is aggregated from the forecasts for its four largest countries (each obtained with a "small" country data set). While trying to include a broad range of specifications, we also aim at not "duplicating" models (including model versions that produce similar results).

The probabilistic forecasts from individual models are combined via a linear prediction pool with optimal weights, chosen so as to maximise forecast accuracy (Geweke and Amisano, 2011). Forecast combinations have frequently been found in empirical studies to produce better forecasts than methods based on the ex ante best individual forecasting model. They have come to be viewed as a simple and effective way to improve and robustify the forecasting performance of

individual models, which are subject to problems such as model misspecification, instability (non-stationarities), and estimation error. This applies to a combination of point forecasts, but also to a combination of probability forecasts.

In the following sections, we present in more details the individual types of models, the data used, the combination method, and the forecast performance thereof.

2.1 Bayesian VAR types

The following model types are included in the optimal pool:

1. VAR with "Minnesota" priors

$$Y_t = c + \sum_{i=1}^{p} B_i Y_{t-i} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, H_t), \tag{1}$$

where Y_t is the vector of dependent variables, c is the intercept, $B_1 \dots B_p$ are matrices of lagged coefficients and ε_t is a vector of innovations. We consider two versions of the model: (i) "in differences" (*Minn Dif*) in which the trending variables are transformed as log-differences and (ii) "in levels" (*Minn Lev*) in which those variables are taken in logs.

We use independent normal priors for the coefficients B_i. The prior means are equal to 0, with the exception of the prior for the diagonal of B_1 (first lag of the dependent variable in each equation) for the specification "in levels", which is equal to 1. Following the "Minnesota" convention, the coefficients for more distant lags are "shrunk" more (have tighter priors around 0). The priors' variances are also adjusted for relative differences in predictability. The overall degree of shrinkage, as governed by the hyperparameter λ, is set to the standard value of 0.2.[1] The prior for the intercept c is non-informative. See Kadiyala and Karlsson (1997), Bańbura et al. (2010), and Carriero et al. (2019) for more details on this type of models, which represent one of the most popular implementation of Bayesian VARs in recent applications in macroeconomics.

2. VAR with democratic priors (*Dem*)

$$Y_t = \mu + \sum_{i=1}^{p} B_i (Y_{t-i} - \mu) + \varepsilon_t, \qquad \varepsilon_t \sim N(0, H_t). \tag{2}$$

In contrast to the previous version, this parameterisation of the VAR model is in deviation from the unconditional mean μ (sometimes referred to as the "steady state"). Informative normal priors are used for μ (Villani, 2009) and stochastic volatility for ε_t (Clark, 2011). As the mean of the prior we take the long-term forecasts from Consensus Economics (as in the "democratic prior" approach proposed by Wright, 2013).[2] As this type of parameterisation

[1] We also evaluated the hierarchical approach of Giannone et al. (2015); the performance was similar to the implementation with fixed λ.

[2] For the variables for which the long-term survey forecasts (of sufficient length) are not available (e.g. interest rates), we use non-informative priors.

assumes the existence of constant unconditional mean, it is only used for variables "in differences". The priors for B_i are the same as above.

3. VAR with local mean (LM)

$$Y_t - \mu_t = \sum_{i=1}^{p} B_i \left(Y_{t-i} - \mu_{t-i}\right) + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, H_t\right), \tag{3}$$

$$\mu_t = \mu_{t-1} + \eta_t, \qquad \eta_t \sim N(0, V_t). \tag{4}$$

The VAR is written in deviation from a "local mean", μ_t, that can vary over time as a random walk (reflecting e.g. low frequency changes in demographics, productivity, or inflation trend/expectations). For reasons similar to above, the specification is meaningful for variables "in differences". The priors for B_i are the same as above.

4. VAR with local mean linked to long-term expectations (SLM)

This type of VAR is specified as the previous one, but in addition the local mean is linked to the long-term forecasts from Consensus Economics, z_t:

$$z_t = \mu_t + g_t, \qquad g_t \sim N\left(0, G_t\right). \tag{5}$$

Again, the specification is meaningful for variables "in differences" and the priors for B_i are the same as above. See Bańbura and van Vlodrop (2018) for implementation details.

5. Time-varying parameters VAR with stochastic volatility (TVP)

$$Y_t = c_t + \sum_{i=1}^{p} B_{i,t} Y_{t-i} + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, \Sigma_t\right), \tag{6}$$

$$c_t = c_{t-1} + \eta_t, \qquad \eta_t \sim N(0, U_t^c), \tag{7}$$

$$B_{i,t} = B_{i,t-1} + \eta_t, \qquad \eta_t \sim N(0, U_t^B). \tag{8}$$

This is the standard implementation of the VAR where all the coefficients can vary over time; see Primiceri (2005) and Del Negro and Primiceri (2015).

6. Univariate unobserved component model with stochastic volatility ($UCSV$)

This is a "non-centred" version of the UCSV model as per Stock and Watson (2007), with gamma priors on the error variances in the two stochastic volatility state equations (see Chan, 2018). The model decomposes each variable into a trend and a transitory component, where each component features an independent stochastic volatility:

$$y_t = \tau_t + e^{\frac{1}{2}(h_0 + \omega_h \tilde{h}_t)} \varepsilon_t^y, \qquad \varepsilon_t^y \sim N(0,1), \tag{9}$$

$$\tau_t = \tau_{t-1} + e^{\frac{1}{2}(g_0 + \omega_g \tilde{g}_t)} \varepsilon_t^\tau, \qquad \varepsilon_t^\tau \sim N(0,1), \tag{10}$$

$$\tilde{h}_t = \tilde{h}_{t-1} + \varepsilon_t^h, \qquad \varepsilon_t^h \sim N(0,1), \tag{11}$$

$$\tilde{g}_t = \tilde{g}_{t-1} + \varepsilon_t^g, \qquad \varepsilon_t^g \sim N(0,1). \tag{12}$$

This is the only model estimated variable by variable, therefore y_t is a scalar.

Regarding the time-varying variances of the VAR innovations, H_t, in models 1-4 and 6 we follow the approach for stochastic volatility of Carriero et al. (2019).[3]

Specifications 2. and 4. allow including off-model information that helps to pin down the unconditional mean and the trend, respectively. For this we use long-term forecasts from Consensus Economics. Alternative sources of information could be considered as well, notably estimates of potential growth to provide information on likely low frequency evolution of real GDP growth (and possibly also of expenditure components).

2.2 Data set compositions

We build a real-time data set in order to simulate the environment available to SPF respondents and policy makers at the time of the survey rounds. The data set is built from the historical vintages from the ECB's Statistical Data Warehouse (SDW), with cut-off dates adjusted to correspond to those of the ECB SPF rounds. Where vintages are not available, we use a pseudo-real time approach, assuming no revisions. For HICP inflation's, vintages before 2006 were not seasonally adjusted, therefore we adjust them using X11. Most series were backdated to 1970 using the Area Wide Model (AWM) database (Fagan et al., 2005). The sample for the euro area goes from 1970Q1 to 2019Q4. The first ten years of data (up to 1980Q1) are used as training sample. Estimation starts in 1980Q2 and forecasts are done for the period 2001Q1 - 2019Q4.

The model variants also differ in terms of data set composition. For the models with aggregate euro area data we use data sets with three and 19 variables. The latter is not feasible for some model types due to computational reasons. The model specifications are summarised in Table 1. A detailed list of variables along with the applied transformations is provided in Table A.1 in the Appendix.

We also obtain bottom-up forecasts by aggregating results for the largest four countries of the euro area (Germany, France, Italy and Spain). The latter models include country GDP and HICP inflation, as well as the euro area level short-term interest rate. We use the period 1980Q1 - 1985Q4 as training sample. Estimation sample starts in 1986Q1. The forecasts are done for the same periods as for the euro area data.

Table 1: Data set compositions

	Minn Dif	Minn Lev	Dem	LM	SLM	TVP	UCSV
Euro area, 3 variables	x	x	x	x	x	x	x
Euro area, 19 variables	x	-	x	-	-	-	-
Big 4, 3 variables	x	-	x	-	x	x	-

[3]The matrix of impulse response functions is assumed constant and log-variances follow univariate random walks. B_is are sampled equation by equation as in Carriero et al. (2021) to speed up the computations.

2.3 Combination method

We combine predictive densities from individual models via a linear optimal pool, where each model contributes to the combination with a time-varying weight driven by the model's performance in terms of predictive density (see Geweke and Amisano, 2011), namely according to the log scoring criterion:

$$\sum_{s=T_1}^{t} log\left(p(y_s; Y_{s-h}, \ldots, Y_1, M)\right), \tag{13}$$

where $p(y_s; Y_{s-h}, \ldots, Y_1, M)$ is the predictive density from model M for variable y_s given the data $Y_1, \ldots, Y_{s-h}$, approximated using a non-parametric kernel smoother. The individual predictive densities are obtained by simulating the parameters from the posterior distribution and drawing the "future" shocks. The optimal weights are found by solving the following constrained maximisation problem:

$$w^*_{t+h|t} = \arg\max_{w_i} \sum_{s=T_1}^{t} \log\left[\sum_{i=1}^{I} w_i p(y_s; Y_{s-h}, \ldots, Y_1, M_i)\right], \tag{14}$$

where I is the number of models, $w^*_{t+h|t} = (w^*_{t+h|t,1}, \ldots, w^*_{t+h|t,I})$ and $w^*_{t+h|t,i}$ is the time-varying weight for model M_i.[4] The weights are constrained to be non-negative and sum to one:

$$w^*_{t+h|t,i} \geq 0, \qquad \sum_{i=1}^{I} w^*_{t+h|t,i} = 1.$$

The combined density is a mixture of the individual densities, weighted over time by the optimal weights found in (14). For a discussion on proper scoring rules for optimal pooling, see for example Martin et al. (2021).

2.4 Performance of Optimal Pooling

We look at each individual model and at their combination's performance, according to the following dimensions, described below: relative point accuracy (RMSFE), relative density accuracy (LPS and CRPS), absolute accuracy (PITs), and time-varying relative accuracy (cumulative RMSFE, LPS, and CRPS with respect to SPF forecasts).

The root mean squared forecast error (*RMSFE*):

$$\frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} (y_t - \hat{y}^i_{t|t-h})^2,$$

[4]Note that for the first 8 quarters of the evaluation sample we assume equal weights and the optimisation is done for $t = T_1 + 8, \ldots, T_2$.

where T_1 and T_2 denote the beginning and the end of the evaluation sample, respectively, and $\hat{y}^i_{t|t-h}$ denotes the median of the predictive density for y_t given the data up to $t-h$ for model (or combination) i. The measure is used to compare the accuracy of point forecasts.

The continuous ranked probability score ($CRPS$):

$$\frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} \left(\int_{-\infty}^{\infty} (F(y; Y_{t-h}, \ldots, Y_1, M_i) - I(y_t \leq y))^2 dy \right) ,$$

where $F(\cdot; Y_{t-h}, \ldots, Y_1, M_i)$ denotes the predictive cumulative distribution function (corresponding to the predictive density $p(\cdot; Y_{t-h}, \ldots, Y_1, M_i)$) and $I(\cdot)$ is an indicator function. The CRPS (Gneiting and Raftery, 2007) considers the cumulative density of the forecast and its distance from the realisation; the lower the score, the better the model accuracy.

The log-predictive score (LPS):

$$\frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} \log(p(y_t; Y_{t-h}, \ldots, Y_1, M_i)) ,$$

where $p(y_t; Y_{t-h}, \ldots, Y_1, M_i)$ is a predictive density for y_t using information up to time $t-h$. The log score is heavily affected by more extreme realisations. It may be therefore necessary to replace particularly low values of predictive densities by a normalising constant (close to zero) in order to avoid the log score reaching minus infinity.

The probability integral transform (PIT):

$$PIT_t = F(y_t; Y_{t-h}, \ldots, Y_1, M_i), \qquad t = T_1, \ldots, T_2 ,$$

provides a measure of the model calibration; for well-calibrated predictive distribution (i.e. such that approximates well the actual distribution), the sequence $PIT_{T_1}, \ldots, PIT_{T_2}$ should be uniformly distributed over the interval $[0, 1]$. To test the hypothesis of uniformity, we perform the Berkowitz test (Berkowitz, 2001)[5].

Cumulative relative RMSFE, CRPS, and LPS: we take the difference of each model combination's score with respect to the corresponding SPF's score, and cumulate it over time.

Results for the individual models are presented in Appendix B. We also check the performance of a combination with equal weights, $w_i = \frac{1}{I}$, for robustness. We find that equal weights improve on most individual models, albeit not as much as the optimal pool. Only a few individual models do better than the optimal pool, and none of them does better in all metrics, variables and horizons. For inflation, the pooling seems to add particular value, with gains from individual models in the order of $10 - 20\%$.

[5] We also try the Knüppel test for the calibration of multi-step ahead density forecasts (Knüppel, 2015), obtaining similar qualitative, albeit less discriminating, results. Omitted here and available upon request to the authors.

Results from the univariate optimal pool are shown in the first column of Table 2 and in each first panel of Figures 1 - 4, where the first figure shows the forecast distribution with median, 5^{th}, 25^{th}, 75^{th}, and 95^{th} percentile and the second figure shows the probability integral transforms (PITs), measuring calibration. The same figures for each individual model are in Appendix B.

The two panels of Figure 5 investigate the relative performance of each combination method with respect to optimal pooling. CRPS and -LPS are in differences from the optimal pool and cumulated over time. Whenever a line is below zero, the loss function is smaller for the alternative combination than for optimal pooling, therefore optimal pooling's performance is worse. A line above zero indicates that optimal pooling is to be preferred to the alternative. Whenever the line is around zero, the two models have performed similarly (as is the case for the optimal pool with SPF for one-year-ahead GDP forecasts). Figures C.1 - C.5 in the Appendix present analogous results for the two-year-ahead forecasts.

3 Continuous forecast density from discrete SPF histograms

The Survey of Professional Forecasters (SPF) for the European Union has been taking place quarterly since the beginning of 1999. The survey asks to a panel of professional forecasters within the EU to give an estimate on the future values for euro area gross domestic product growth, HICP inflation, and unemployment rate (de Vincent-Humphreys et al., 2019; Kenny et al., 2013). We focus here on GDP and inflation forecasts, for two separate medium-term horizons, namely one- and two-year-ahead[6]. Respondents are asked to give both a point forecast and to assign probabilities for each variable's future outcome falling within pre-determined ranges. The individual responses are then aggregated, and a histogram of average probabilities for the economic outlook results. We do not focus on individual responses, following the results in Genre et al. (2013), where the simple average is proven to be the best combination method. Other aggregation methods include optimal pooling like in Conflitti et al. (2015), and a more recent work by Diebold et al. (2020), where the authors propose to build regularised mixtures of individual densities.

We consider, for each release of the SPF, a point forecast and a histogram of probabilities, for the two target variables and the two horizons. From this information, it would be already possible to calculate the SPF point forecast accuracy, standard deviation, and calibration scores, provided one calculates a discrete approximation of the histogram, and assumes that the probability is concentrated in the mid-point of each bin (Kenny et al., 2015). We follow this approach to calculate mean and standard deviation for the tilting, described in the next section. Another approach (Engelberg et al., 2009) is to assume a normal or a beta distribution for the SPF histograms, however, this method presents the obvious shortcoming of possibly misspecifying the distribution.

For this reason, we decide to take a non-parametric approach (see Billio et al., 2013, for an application to prediction of stock market returns) and build a continuous SPF distribution, as well

[6]For each round, the target quarter refers to one (or two) years after the latest official release available at the time of the questionnaire.

as draws from this distribution, using a kernel smoother. The procedure to build the continuous density for each SPF round is as follows: starting from the aggregated discrete probabilities, we simulate N (=1000) random samples from multinomial distributions, each with sample size equal to the average number of respondents times the number of bins[7], and probabilities equal to the SPF probabilities for each bin. For each sample we then draw numbers included in the corresponding bin, making the assumption that the values are distributed uniformly within the bin. For example, if the multinomial density has 25 counts for the bin with bounds [1.95 2.45), we draw 25 numbers from the uniform distribution with those bounds. Finally, we fit N kernel densities over the grid of values corresponding to the bins[8] and take the average of these N densities. We then simulate S (=5000) draws from the average kernel density over the same grid, obtaining a simulated density forecast, which can be compared to densities from other models, or combined with them.

3.1 Performance of SPF

Looking at SPF results, two main features of the forecasts emerge. First, the survey does extremely well in terms of point accuracy (RMSFE), particularly for HICP inflation, where it does better than optimal pooling (see Table B.1 in the Appendix, second column), with the exception of two-year-ahead GDP. Second, in terms of density accuracy, the SPF tends to always be over-confident, delivering narrow distributions and u-shaped PITs (see second panel of Figures 1 - 4 for the one-year horizon and C.1 - C.4 in Appendix for the two-year horizon). This is reflected also in the log-scores, which, as displayed in the second column of Table 2, are always worse than in the case of optimal pooling (the difference of the two scores is always negative, meaning that the SPF log scores are smaller), by a larger margin in the case of GDP.

In terms of time variation, the SPF performance clearly worsens during and after the global financial crisis. The degree of uncertainty, especially for GDP growth, was not captured by SPF respondents; one reason for the large errors, particularly in density terms, is the fact that the pre-determined bins for the GDP's distribution did not reflect the economic developments. In the SPF round of 2008Q3, for the forecast of GDP referring to 2009Q3, respondents did put a large probability in the left-most bin, which nevertheless represented all values below -1.1%[9]. The realisation for year-on-year GDP growth was close to -5% for that quarter. It is not possible to know how the probability assigned to the left-most bin is distributed across values smaller than -1.1%. For equally sized bins, in this particular round, the beginning of the bin would be -1.5%, returning a predictive score of zero. This in turn would result in a value for the log-predictive score of minus infinity. For this reason, we set a lower limit to the log score of -20, therefore selecting this value whenever the log-score is smaller than -20; the same rule is applied to each

[7]Each respondent can give a probability to each bin, therefore we can approximate the number of total answers as the product of respondents and bins. To simplify, we assume a constant number of 80 respondents for each round. The number of respondents does not impact the final simulated draws.

[8]For the extreme bins of inflation, we assume a bin size equal to all the intermediate ones; for those of GDP, we expand the left extreme bin such that it starts at values of -10. This is to allow a more realistic distribution in this bin during crisis periods.

[9]The bins have been expanded since after that SPF round, with ten additional bins added, for values of GDP growth down to -6.5%

individual model and combination.

4 Incorporating model and survey information

In our first set of results, we find strengths and weaknesses in both judgement-based and combination of model-based forecasts. While the former perform particularly well in the point dimensions, the latter can considerably improve the density and calibration. For this reason, we unify these two sources of information and look at whether and how this improves forecast accuracy.

We focus on two additional methods:

1. Including the simulated SPF density described in Section 3 into the BVAR pool and combining individual models + SPF by means of optimal pooling (*Opt Pool w SPF* in the results);
2. Using entropic tilting, including moments from the SPF, in the following four options[10]:
   (a) Tilt to the SPF mean each individual model, then perform optimal pooling (*Mean-tilted ex-ante*);
   (b) Tilt to the SPF mean and variance each individual model, then perform optimal pooling (*Mean and var-tilted ex-ante*);
   (c) Tilt to the SPF mean the optimal pool of combined models (*Mean-tilted ex-post*);
   (d) Tilt to the SPF mean and variance the optimal pool of combined models (*Mean and var-tilted ex-post*);

4.1 Combination of models with SPF densities

As a first method to add survey information to model forecasts, we take the same forecast densities derived from the models described in Section 2 and the SPF density constructed as in Section 3 and combine them by means of a linear optimal pool. The optimal weights are again found by solving the following constrained maximisation problem, analogous to (14):

$$w^*_{t+h|t} = \arg\max_{w_i} \sum_{s=T_1}^{t} \log \left[\sum_{i=1}^{I+1} w_i p(y_s; Y_{s-h}, \ldots, Y_1, M_i) \right], \tag{15}$$

where $w^*_{t+h|t} = (w^*_{t+h|t,1}, \ldots, w^*_{t+h|t,I+1})$, $w^*_{t+h|t,i}$ is the time-varying weight for model M_i or for the SPF density, and I is the number of models, to which we add the SPF density.

[10] Schemes 2(c)-2(d) are also studied in Galvao et al. (2021).

4.2 Entropic tilting

The second method consists of imposing the SPF moments as restrictions onto each individual density before performing optimal pooling, or onto the optimally combined distribution of forecast densities. To do that, we use a relative entropy approach, as seen in Robertson et al. (2005). The procedure consists of re-weighting the draws from the forecast distribution so that it satisfies the required restrictions while being as close as possible to the original distribution. Tilting has been used in the past to produce conditional forecasts (imposing that the path for some variables over the forecast period is equal to some predetermined quantities, such as in Robertson et al., 2005) or to produce forecasts that satisfy economic theory, by imposing moments such as Euler conditions, as seen in Giacomini and Ragusa (2014). The tilting procedure is relatively straightforward. The following exposition is a summary from Robertson et al. (2005), as it corresponds to the methodology that we use.

For each variable, we start with the *(h x k)* matrix that contains the forecast density – that is, k draws for 1 to h quarters ahead. The distribution is a random sample from the forecasting density, so initially we assign the same initial weight to each draw, $\pi_i = \frac{1}{k}, i = 1, ..., k$. The basic idea of tilting is to modify those weights so that the re-weighted distribution, with weights π_i^*, satisfies the restrictions of interest; in this case, those coming from the SPF. However, the weights are found so that the new distribution is "close" to the original one. In order to measure the closeness of both probability distributions, we use the Kullback-Leibler Information Criterion,

$$K(\pi_i^*, \pi_i) = \sum_{i=1}^{k} \pi_i^* \log(\pi_i^*/\pi_i).$$

The new weights are then calculated so that they minimise $K(\pi_i^*, \pi_i)$, subject to the following constraints,

$$\begin{aligned} &\pi_i^* \geq 0\,, \\ &\sum_{i=1}^{k} \pi_i^* = 1\,, \\ &\sum_{i=1}^{k} \pi_i^* g(y_i) = \bar{g}. \end{aligned}$$

The first two constraints are trivial and imply that the new weights should be positive and should sum to 1. The third constraint imposes the restrictions and implies that the expectations of a function of the draws from the forecasting distribution should be equal to a fixed quantity. For example, if $g(y_i) = y_i$, the restriction is put on the mean of the distribution, which we would fix to $\bar{m}$. In our application, where we use restrictions based on the SPF, we consider the standard deviation, in addition to the mean.

Matching the SPF standard deviation is straightforward. Given a variance $\bar{g_2}$, then,

$$g(y_i) = (y_i - \bar{m})^2.$$

Finally, the minimisation problem yields the following solution for the new weights:

$$\pi_i^* = \frac{\pi_i exp(\gamma' g(y_i))}{\sum_{i=1}^k \pi_i exp(\gamma' g(y_i))}.$$

In this case, γ is the multiplier associated with the restrictions, which can be found numerically as:

$$\gamma = \arg\min_{\tilde{\gamma}} \sum_{i=1}^{k} \pi_i exp(\tilde{\gamma}'[g(y_i) - \bar{g}]).$$

4.3 Performance of model combination with SPF

Results from the various combinations mentioned above can be found in Table 2 and in the following Figures. Since we already described the results for optimal pooling without judgement and for SPF forecasts in Sections 2 and 3, we focus here on the five remaining specifications.

The first method implies adding the forecast distribution simulated for the SPF to the pool of BVARs and finding weights which maximise the predictive likelihood. Figures 10 and C.10 show the resulting optimal weights; Figure 8 and C.8, for reference, show the optimal weights from the pool of BVARs only. The weights for SPF densities are never very large, which is also reflected in the relative cumulative scores of Figures 5 and C.5. The dashed line for optimal pool with SPF is almost always near zero for GDP growth, meaning that CRPS and LPS are similar for both optimal pools, including and excluding SPF. For inflation, the weight of SPF is more persistently positive in the second half of the sample, reflecting in an effect on the pool's accuracy. Looking at average scores in the Table, however, the effect seems to vanish over the sample, with the two optimal pools returning very similar scores (third column in Table 2, scores relative to optimal pool without SPF are close to 1 for CRPS and to 0 for the LPS).

The two pools obtained from tilting to the SPF mean improve considerably on the original optimal pool and on the SPF; RMSFE, CRPS, and LPS all get better, especially for inflation. Over time, as shown in the cumulative scores, both options perform very well for GDP at both horizons, as well as for inflation. The dashed lines referring to the mean-tilted combinations are always below the others, particularly since the global financial crisis, when model performance begins to diverge across combinations. The only exception is again the two-year-ahead GDP, where, in terms of CRPS, the optimal pool does better and, in terms of LPS, the optimal pool performs very similarly to the mean-tilted combinations.

The mean-tilted ex-ante pool, in particular, always outperforms the other methods, both on average and over time. This is due to the tilting of individual models, which results in improved initial conditions. Consequently, different models are selected by the optimal pooling, resulting in different combination weights (see Figure 9, compared to the weights for non-tilted models in Figure 8), and better final results. The improved initial conditions are clearly visible in Tables D.1 and D.2 in the Appendix, showing scores for the non-tilted and the mean-tilted individual models relative to the optimal pool, for GDP and HICP, respectively. For example, for GDP 4-quarters ahead the models with the best log score after tilting are the time-varying parameters

(TVP), the local mean (LM) and the Minnesota in levels (MinnL). These are also the models which get a positive weight for the majority of the sample (see Figure 9). The large Minnesota (Minn Large) and the survey local mean multi-country (SLM MC) models, which were obtaining a positive weight before tilting, are now excluded from the pool. For HICP inflation, the TVP and Minnesota multi-country are included in the pool after tilting, albeit with smaller weights, while the SLM is excluded. A larger weight is given to the DPSV and the DP Large models.

Finally, for the case of the tilting to both first *and* second moments of SPF, there is a general worsening of the performance for all variables and horizons. We notice from the distribution and the PITs charts that this method seems to forsake model information and closely replicates the SPF forecast performance. We conclude that it is counterproductive to include too much survey information when this is not well calibrated.

5 Bivariate analysis

In this section, we describe results analogous to Section 4 for a bivariate setting. We perform optimal pooling for both GDP and inflation by maximising the bivariate predictive performance, i.e. the predictive performance of the joint pdf of both variables. This analysis allows to investigate whether models or surveys' performance for one of the two variables or some specific area of interest of the same variable can be explained by conditioning on the other variable.

To measure accuracy, we consider three scoring functions. We look at the energy score, a multivariate generalisation of the CRPS (see Gneiting et al., 2008), the bivariate LPS, and the conditional PITs (Diebold et al., 1999).

The energy score (ES) is defined as:

$$\frac{1}{S}\sum_{i=1}^{S}\|\tilde{Y}_i - Y\| - \frac{1}{2S^2}\sum_{i=1}^{S}\sum_{j=1}^{S}\|\tilde{Y}_i - \tilde{Y}_j\|$$

where S is the number of draws of forecast densities for each period, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$ (here, $d = 2$), $\tilde{Y}$ is the vector of forecast densities, and Y is the vector of realisations.

The bivariate log-predictive score (LPS),

$$\frac{1}{T_2 - T_1 + 1}\sum_{t=T_1}^{T_2}\log(p(Y_t|Y_{t-h},\ldots,Y_1))$$

is analogous to the univariate case, but here Y_t is a vector and not a scalar.

The conditional PIT is given by:

$$PIT_t(y_{1t} \mid y_{2t}) = F(y_{1t} \mid y_{1t-h},\ldots,y_{11},y_{2t-h},\ldots,y_{21}), \qquad t = T_1,\ldots,T_2$$

This is obtained by first calculating the univariate PITs (based on the CDF of the first variable), then the conditional PITs (PITs of the second variable, given that the PITs of the first variable fall within chosen quantiles). The procedure for the bivariate evaluation results in two conditional PITs, $PIT_t(GDP \mid HICP)$ and $PIT_t(HICP \mid GDP)$, respectively. In our results, we focus on two parts of the PITs distribution; a "central" part, given by the central 68% of the distribution and a "tails" part, which is the sum of quantiles $0^{th} - 16^{th}$ and $84^{th} - 100^{th}$. The conditional PITs are evaluated in the same way as the unconditional ones; for well-calibrated predictive distributions (i.e. such that approximates well the actual distribution), the sequence $PIT_{T_1}, \ldots, PIT_{T_2}$ should be uniformly distributed over the interval $[0, 1]$. To test the hypothesis of uniformity, we again perform the Berkowitz test (Berkowitz, 2001).

5.1 Performance of bivariate combinations

Table 3 shows results for the bivariate combinations. As for the univariate case, ES and LPS of the different forecasts are shown relative to the first column, optimal pooling. The Berkowitz test is in absolute terms. Results are consistent with the univariate case. According to the LPS metric, the optimal pool tilted ex-ante to the mean of SPF does best, followed by the optimal pool tilted ex-post to the mean in the one-year-ahead and by the optimal pool with SPF in the two-year-ahead. The ES, similarly to the CRPS, penalises less the SPF and the combinations tilted to both mean and variance of SPF. At the two-year horizon, only the pooling with SPF does better than optimal pooling for this metric.

Conditional calibration can be assessed both in Table 3 and in Figures 6-7. The latter focuses only on the best and the worst performing combinations and shows both the unconditional and the conditional PITs for each variable for the two combinations. The two subplots in the bottom part of the Figure include conditional PITs for the "center" and the "tails". For example, the middle subplot of Figure 6 shows the PITs of GDP conditional on the PITs of inflation falling in the mid-68% of the distribution at one-year-ahead horizon. The bottom subplot shows the PITs of GDP, conditional on the PITs of inflation falling either in the first 16% or in the last 84% of the distribution. The mean-tilted ex-ante combination satisfies uniformity for almost all conditional PITs, failing only for the two-year-ahead PITs of GDP conditional on the extremes ($< 16\%$ and $> 84\%$) of inflation. The mean and variance-tilted combination, in contrast, carries over the poor unconditional calibration also in the conditional case. Table 3 indicates that the SPF is not correctly calibrated in any case. The lack of calibration of the SPF is not documented by the evidence that the survey respondents give higher relevance to one of the two variables, but their densities are simply too narrow.

Finally, we look at the weights from optimal pooling for the bivariate case, in Figures 8 (pool without SPF) and 10 (pool including SPF). There are large differences in terms of selected models compared to the univariate optimal pool, which include SPF getting a larger weight both before and after the 2008 crisis, new models entering the pool with positive weight, and other models being excluded (notably the SLM model, which for inflation had a large positive weight during most of the sample). Weights in the multivariate case are less volatile than those for univariate combinations of GDP, in particular at the beginning of the sample, but still more volatile than

those for univariate combinations of inflation. Despite these differences in weights visible also in the two-year-ahead results in the Appendix (Figures C.8-C.10), results in terms of density performance are very similar to the univariate pooling, as discussed above. The main difference to notice is the improvement in GDP point forecasting accuracy when applying bivariate combinations (see Table B.1): five cases over the six combination schemes have lower root mean square forecasting errors in the bivariate case than in the univariate case. However, the evidence is opposite for inflation forecasting, where performance deteriorates in the bivariate case.

6 Conclusions

We evaluate in real time density forecasts from a broad range of Bayesian VARs for euro area GDP and inflation. We look at average density performance over the sample, at overall calibration, and at relative performance over time. We then combine results from individual models by means of a linear optimal pool, finding significant improvements with respect to each model and to a trivial combination with equal weights. In addition, we consider forecasts from the Survey of Professional Forecasters and build a continuous distribution and simulated draws from the survey's aggregate histograms. Evaluating these forecasts along the same dimensions, we see a very good performance in terms of point forecast, but a poor performance for density and calibration, with SPF density forecasts being overconfident and poorly calibrated. In order to exploit the positive point performance, we include complete distributions for SPF forecasts in the model densities via two methods. First, we add the SPF as an additional density forecast to the linear optimal pool. Gains with respect to the original optimal pool in this case are limited. Second, we use first and second moments from the SPF's histograms to tilt the model densities. We tilt both the individual model densities before combining them and the already combined density. We find that when both moments are used for the tilting, there is a general worsening of the performance, both for the models tilted ex-ante (before combination) and ex-post (after combination). In the case of tilting only to the first moment, all results improve with respect to other alternatives and with respect to SPF only, particularly when individual models are tilted before being combined. Evidence is similar when the evaluation is carried out jointly for the two variables and not marginally as is commonly reported in literature. Therefore, the lack of calibration of the SPF cannot be justified by the survey respondents giving higher relevance to one of the two variables, but by the evidence that survey densities are simply too narrow.

We conclude that there are some benefits to including forward-looking medium-term judgement forecasts to a combination of purely statistical models, with improvements in the point forecast accuracy which are not achieved by a simple optimal pooling of those models. Good forecast calibration and density performance, on the other hand, come mostly from combining individual models, confirming the advantages of hedging against model uncertainty by using several specifications.

References

Altavilla, C., Giacomini, R., and Ragusa, G. (2017). Anchoring the yield curve using survey expectations. *Journal of Applied Econometrics*, 32(6):1055–1068.

Amisano, G. and Geweke, J. (2017). Prediction using several macroeconomic models. *The Review of Economics and Statistics*, 99(5):912–925.

Ang, A., Bekaert, G., and Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54(4):1163–1212.

Angelini, E., Lalik, M., Lenza, M., and Paredes, J. (2019). Mind the gap: A multi-country bvar benchmark for the eurosystem projections. *International Journal of Forecasting*, 35(4).

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

Bańbura, M. and van Vlodrop, A. (2018). Forecasting with Bayesian vector autoregressions with time variation in the mean. Tinbergen Institute Discussion Papers 18-025/IV, Tinbergen Institute.

Bassetti, F., Casarin, R., and Ravazzolo, F. (2020). Density forecasting. In *Macroeconomic Forecasting in the Era of Big Data*, pages 465–494. Springer.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177:213–232.

Carriero, A., Chan, J., Clark, T. E., and Marcellino, M. (2021). Corrigendum to: Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors.

Carriero, A., Clark, T. E., and Marcellino, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212(1):137 – 154. Big Data in Dynamic Predictive Econometric Modeling.

Chan, J. C. (2018). Specification tests for time-varying parameter models with stochastic volatility. *Econometric Reviews*, 37(8):807–823.

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. (2014). Forecast uncertainty—ex ante and ex post: Us inflation and output growth. *Journal of Business & Economic Statistics*, 32(2):206–216.

Clements, M. P. (2018). Are macroeconomic density forecasts informative? *International Journal of Forecasting*, 34(2):181–198.

Conflitti, C., De Mol, C., and Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31(4):1096–1103.

Crump, R. K., Eusepi, S., Giannone, D., Qian, E., and Sbordone, A. M. (2021). A Large Bayesian VAR of the United States Economy. Staff Reports 976, Federal Reserve Bank of New York.

Crump, R. K., Eusepi, S., and Moench, E. (2016). The term structure of expectations and bond yields. Staff Reports 775, Federal Reserve Bank of New York.

de Vincent-Humphreys, R., Dimitrova, I., Falck, E., and Henkel, L. (2019). Twenty years of the ECB Survey of Professional Forecasters. *Economic Bulletin Articles*, 1.

Del Negro, M., Giannone, D., Giannoni, M. P., and Tambalotti, A. (2017). Safety, Liquidity, and the Natural Rate of Interest. *Brookings Papers on Economic Activity*, 48(1 (Spring):235–316.

Del Negro, M. and Primiceri, G. E. (2015). Time varying structural vector autoregressions and monetary policy: a corrigendum. *The Review of Economic Studies*, 82(4):1342–1345.

Diebold, F. X., Hahn, J., and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics*, 81(4):661–673.

Diebold, F. X., Shin, M., and Zhang, B. (2020). On the aggregation of probability assessments: Regularized mixtures of predictive densities for eurozone inflation and real interest rates. *arXiv preprint arXiv:2012.11649*.

Domit, S., Monti, F., and Sokol, A. (2016). A Bayesian VAR benchmark for COMPASS. Bank of England working papers 583, Bank of England.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, 27(1):30–41.

Fagan, G., Henry, J., and Mestre, R. (2005). An area-wide model for the euro area. *Economic Modelling*, 22(1):39–59.

Galvao, A. B., Garratt, A., and Mitchell, J. (2021). Does judgment improve macroeconomic density forecasts? *International Journal of Forecasting*. Forthcoming.

Ganics, G. and Odendahl, F. (2021). Bayesian VAR forecasts, survey information, and structural change in the euro area. *International Journal of Forecasting*, 37(2):971–999.

Garnier, C., Mertens, E., and Nelson, E. (2015). Trend Inflation in Advanced Economies. *International Journal of Central Banking*, 11(S1):65–136.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141.

Giacomini, R. and Ragusa, G. (2014). Theory-coherent forecasting. *Journal of Econometrics*, 182(1):145–155.

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

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.

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23(1):1–13.

Jore, A. S., Mitchell, J., and Vahey, S. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, 25(4):621–634.

Kadiyala, K. R. and Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, pages 99–132.

Kenny, G., Kostka, T., and Masera, F. (2013). Can macroeconomists forecast risk? event-based evidence from the euro area SPF. *ECB Working Paper*.

Kenny, G., Kostka, T., and Masera, F. (2014). How informative are the subjective density forecasts of macroeconomists? *Journal of Forecasting*, 33(3):163–185.

Kenny, G., Kostka, T., and Masera, F. (2015). Density characteristics and density forecast performance: a panel analysis. *Empirical Economics*, 48(3):1203–1231.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33(2):270–281.

Krüger, F., Clark, T. E., and Ravazzolo, F. (2017). Using entropic tilting to combine BVAR forecasts with external nowcasts. *Journal of Business & Economic Statistics*, 35(3):470–485.

Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., and Ramírez-Hassan, A. (2021). Optimal probabilistic forecasts: When do they work? *International Journal of Forecasting*.

Mertens, E. (2016). Measuring the level and uncertainty of trend inflation. *The Review of Economics and Statistics*, 98(5):950–967.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Robertson, J. C., Tallman, E. W., and Whiteman, C. H. (2005). Forecasting using relative entropy. *Journal of Money, Credit and Banking*, pages 383–401.

Sims, C. A. and Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, 39(4):949–68.

Stock, J. H. and Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and banking*, 39(s1):3–33.

Tallman, E. W. and Zaman, S. (2020). Combining survey long-run forecasts and nowcasts with bvar forecasts using relative entropy. *International Journal of Forecasting*, 36(2):373–398.

Villani, M. (2009). Steady-state priors for vector autoregressions. *Journal of Applied Econometrics*, 24(4):630–650.

Wright, J. H. (2013). Evaluating real-time VAR forecasts with an informative democratic prior. *Journal of Applied Econometrics*, 28(5):762–776.

Table 2: Relative accuracy scores and uniformity test results for the different univariate combinations:

	Optimal Pool: *absolute scores*	SPF	Optimal Pool with SPF	μ-tilted ex-ante	μ-tilted ex-post	μ and σ-tilted ex-ante	μ and σ-tilted ex-post
GDP 4-q							
CRPS	0.808	**0.994**	**0.997**	**0.935**	**0.932**	**0.966**	**0.971**
LPS	-1.922	-0.627	**0.030**	**0.302**	**0.026**	-0.406	-0.485
Berkowitz	0.042	0.000	0.016	**0.624**	**0.279**	0.000	0.000
GDP 8-q							
CRPS	0.994	1.091	1.001	1.080	1.033	1.102	1.099
LPS	-1.973	-1.112	-0.094	-0.042	-0.095	-1.243	-1.303
Berkowitz	0.020	0.000	0.011	**0.099**	0.004	0.000	0.000
HICP 4-q							
CRPS	0.503	**0.932**	**0.991**	**0.917**	**0.937**	**0.943**	**0.944**
LPS	-1.306	-0.024	**0.003**	**0.117**	**0.056**	-0.007	-0.082
Berkowitz	**0.839**	0.002	**0.704**	**0.218**	**0.156**	0.000	0.000
HICP 8-q							
CRPS	0.567	**0.949**	1.020	**0.922**	**0.941**	**0.964**	**0.963**
LPS	-1.429	-0.040	-0.001	**0.082**	**0.032**	-0.263	-0.284
Berkowitz	**0.552**	0.000	**0.961**	**0.368**	**0.232**	0.000	0.000

Note: CRPS is calculated as the ratio between each model's score and those of optimal pooling, included in column 1. A number smaller than one indicates a preference for the forecast in that column over optimal pooling. LPS is calculated as the difference between the two scores, therefore a positive value indicates a preference for the forecast over optimal pooling. Berkowitz test is in absolute terms, where a p-value smaller than 0.10 indicates that the null hypothesis of good calibration can be rejected at the 10 percent confidence level; i.e. the density is not well calibrated.

Table 3: Relative accuracy scores and uniformity test results for the different bivariate combinations:

		Optimal Pool: *absolute scores*	SPF	Optimal Pool with SPF	μ-tilted ex-ante	μ-tilted ex-post	μ and σ-tilted ex-ante	μ and σ-tilted ex-post
4-q								
ES		1.015	**0.995**	**0.987**	**0.961**	**0.954**	**0.983**	**0.984**
LPS		-3.355	-0.416	**0.144**	**0.433**	**0.220**	-0.106	-0.352
Berkowitz	c	**0.538**	0.000	**0.907**	**0.690**	**0.650**	0.000	0.000
$y \mid h$	t	0.008	0.000	0.010	**0.112**	0.098	0.000	0.000
Berkowitz	c	0.026	0.013	0.078	**0.219**	**0.168**	0.002	0.001
$h \mid y$	t	**0.141**	0.061	**0.235**	**0.394**	**0.724**	0.010	0.003
8-q								
ES		1.254	1.054	**0.989**	1.024	1.016	1.069	1.068
LPS		-3.766	-0.501	**0.061**	**0.233**	-0.017	-0.782	-0.874
Berkowitz	c	**0.214**	0.000	**0.623**	**0.721**	**0.185**	0.000	0.000
$y \mid h$	t	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Berkowitz	c	0.075	0.002	**0.156**	**0.707**	**0.225**	0.000	0.000
$h \mid y$	t	**0.632**	0.030	**0.915**	**0.238**	**0.383**	0.003	0.002

Note: ES is calculated as the ratio between each model's score and those of optimal pooling. LPS is calculated as the difference between the two scores. Berkowitz test is in absolute terms and refers to the conditional PITs described in Section 5: PITs of GDP given inflation, central part (c) and tails (t); PITs of inflation given GDP, central part and tails. A p-value smaller than 0.10 indicates that the null hypothesis of good calibration can be rejected at the 10 percent confidence level; i.e. the density is not well calibrated.

Figure 1: Densities of one-year-ahead forecasts from combinations and SPF, real GDP growth.

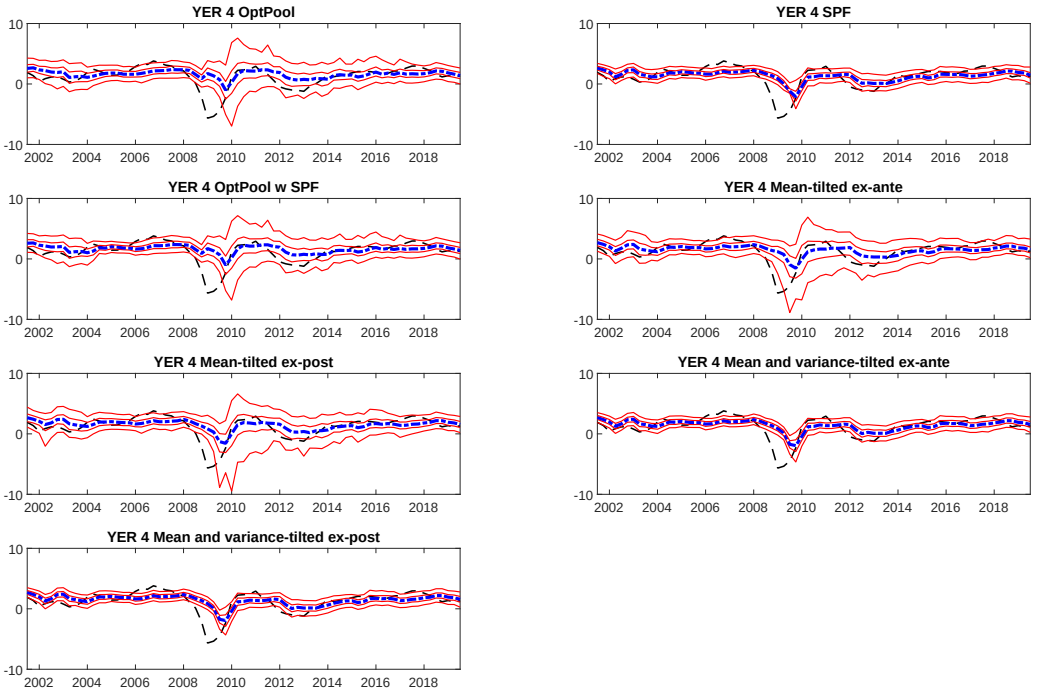


Figure 2: PITs of one-year-ahead forecasts from combinations and SPF, real GDP growth.

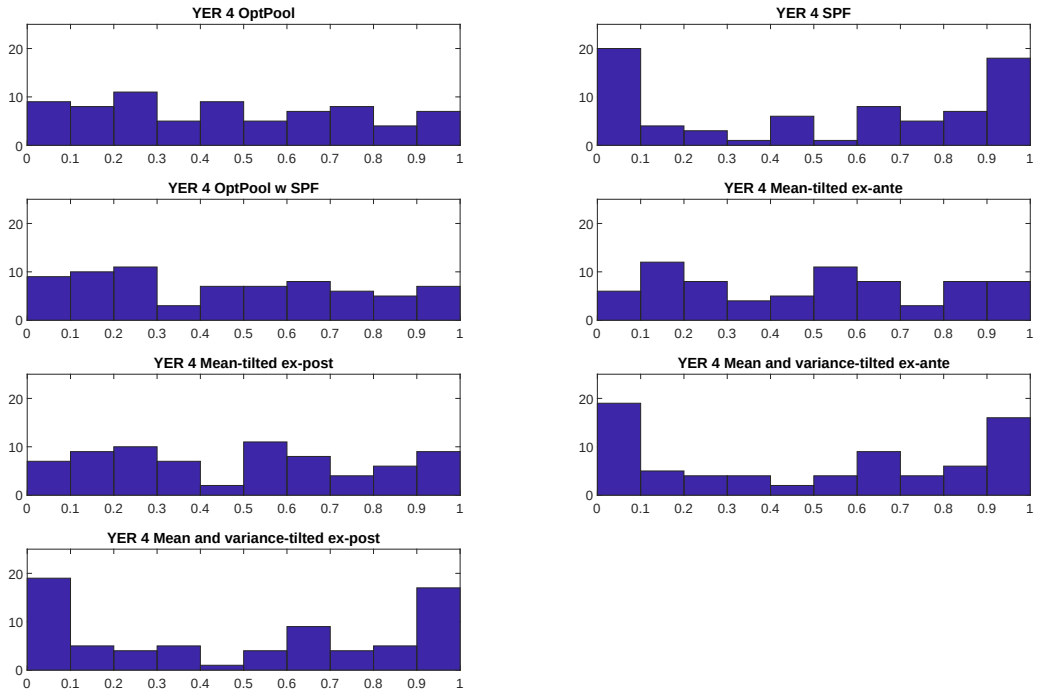


Figure 3: Densities of one-year-ahead forecasts from combinations and SPF, HICP inflation.

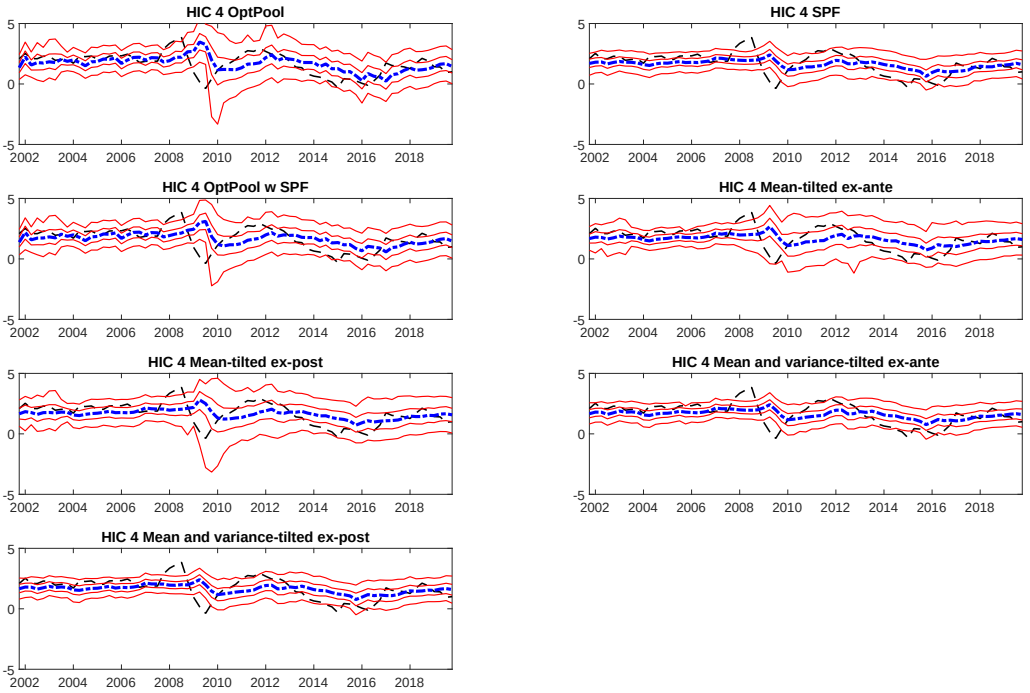


Figure 4: PITs of one-year-ahead forecasts from combinations and SPF, HICP inflation.

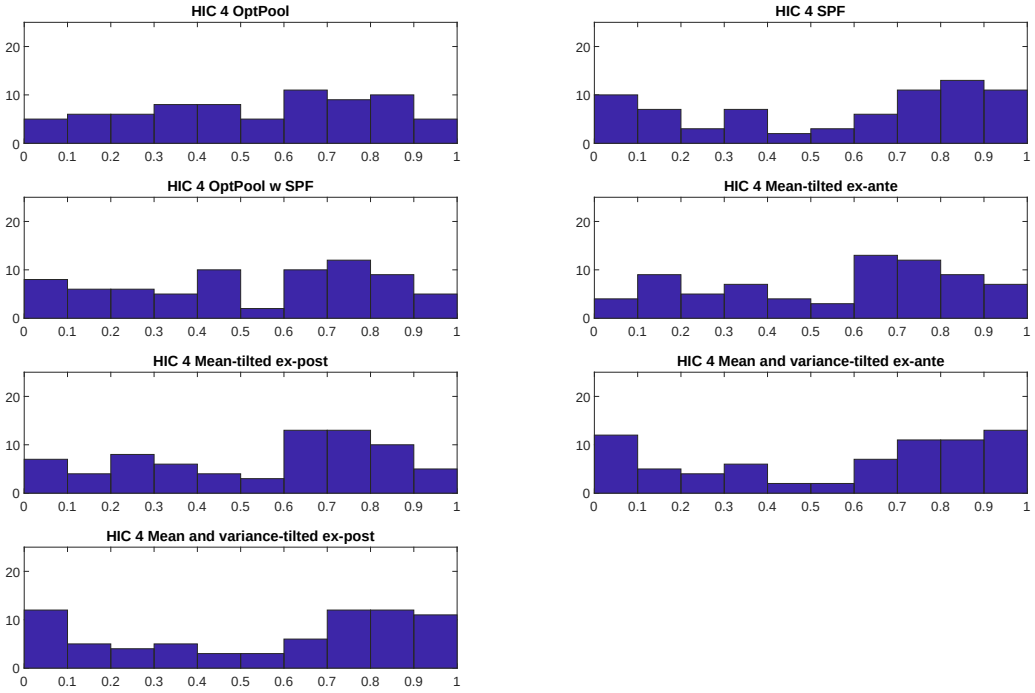


Figure 5: Cumulative relative scores of one-year-ahead forecasts, real GDP growth (left) and HICP inflation (right).

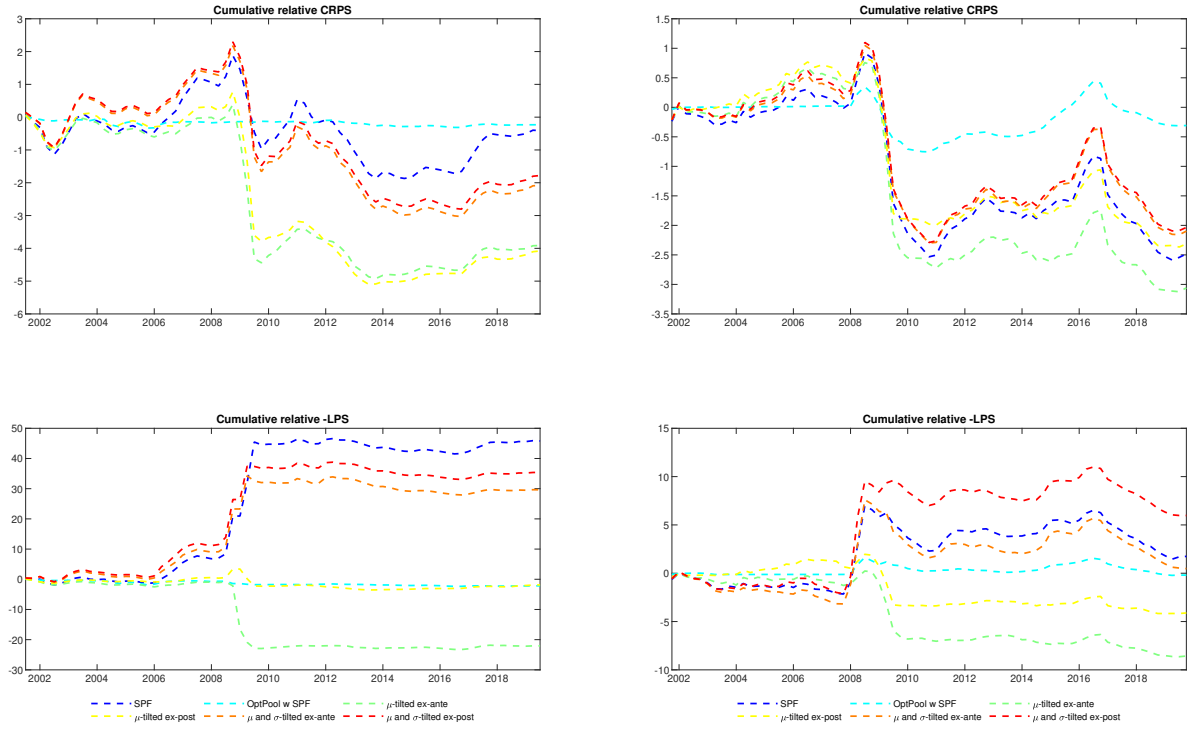


Figure 6: Unconditional and conditional PITs for one-year-ahead real GDP growth, for two combination methods.

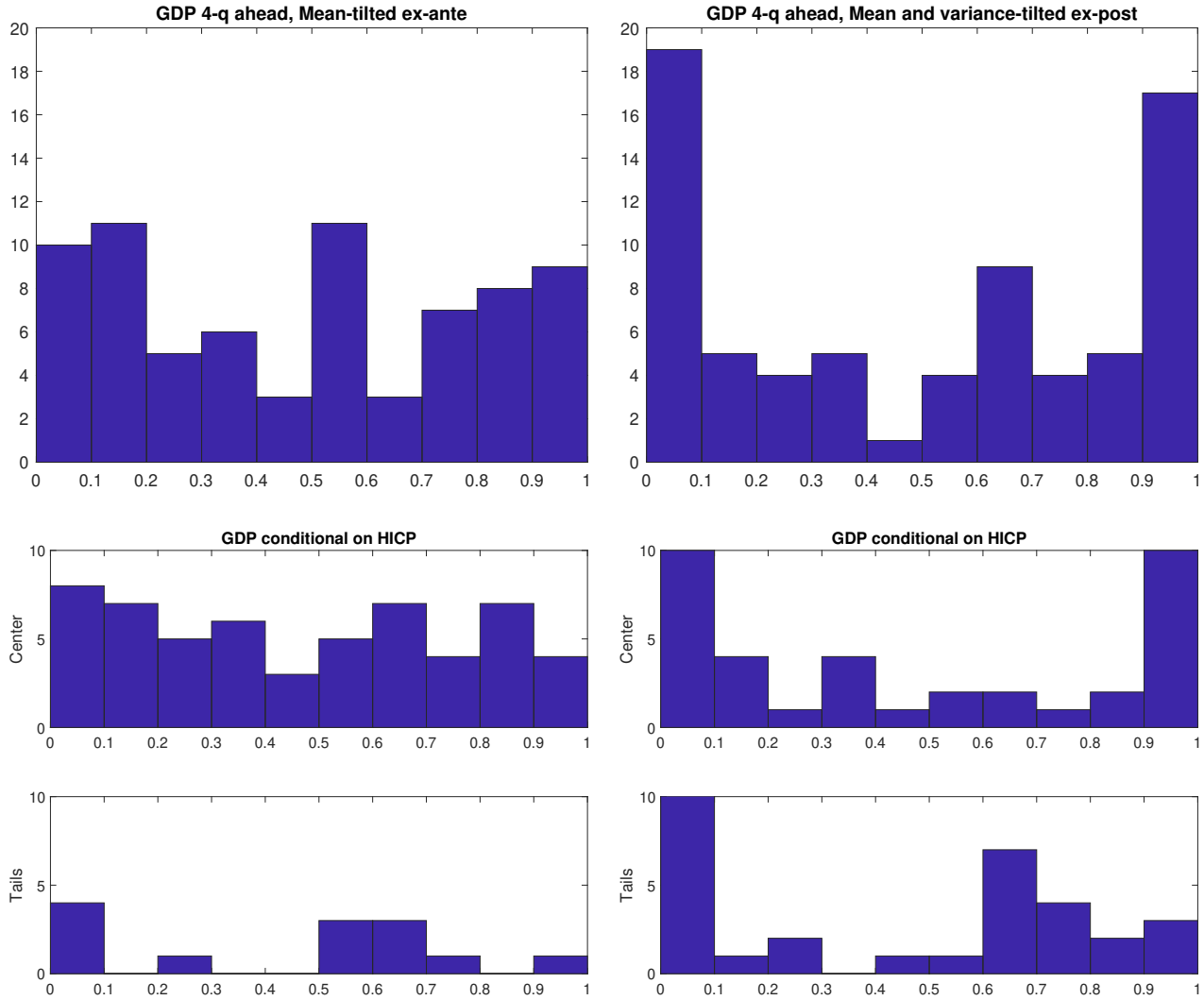


Figure 7: Unconditional and conditional PITs for one-year-ahead HICP inflation, for two combination methods.

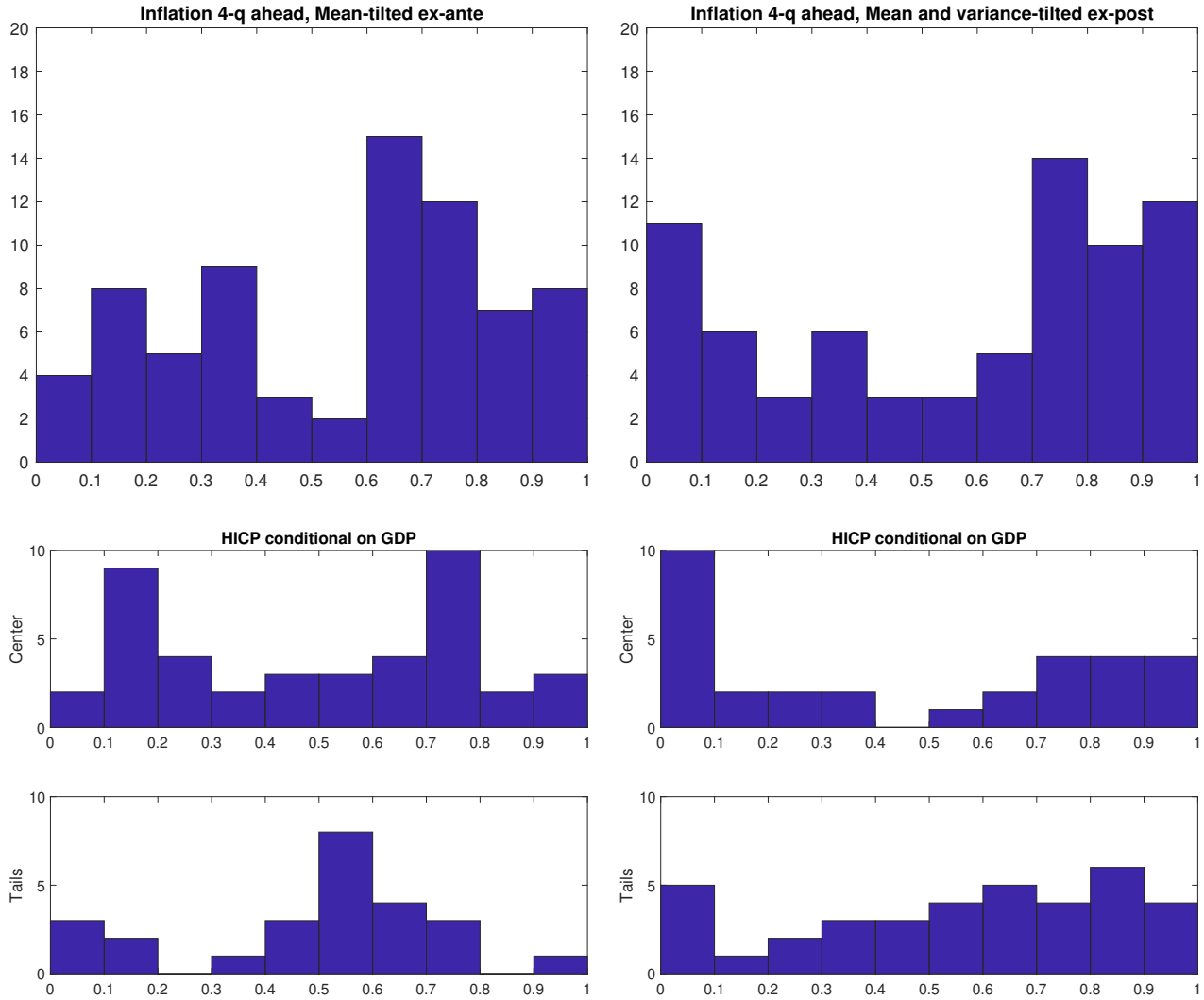


Figure 8: Weights from univariate and bivariate optimal pools, one-year-ahead.

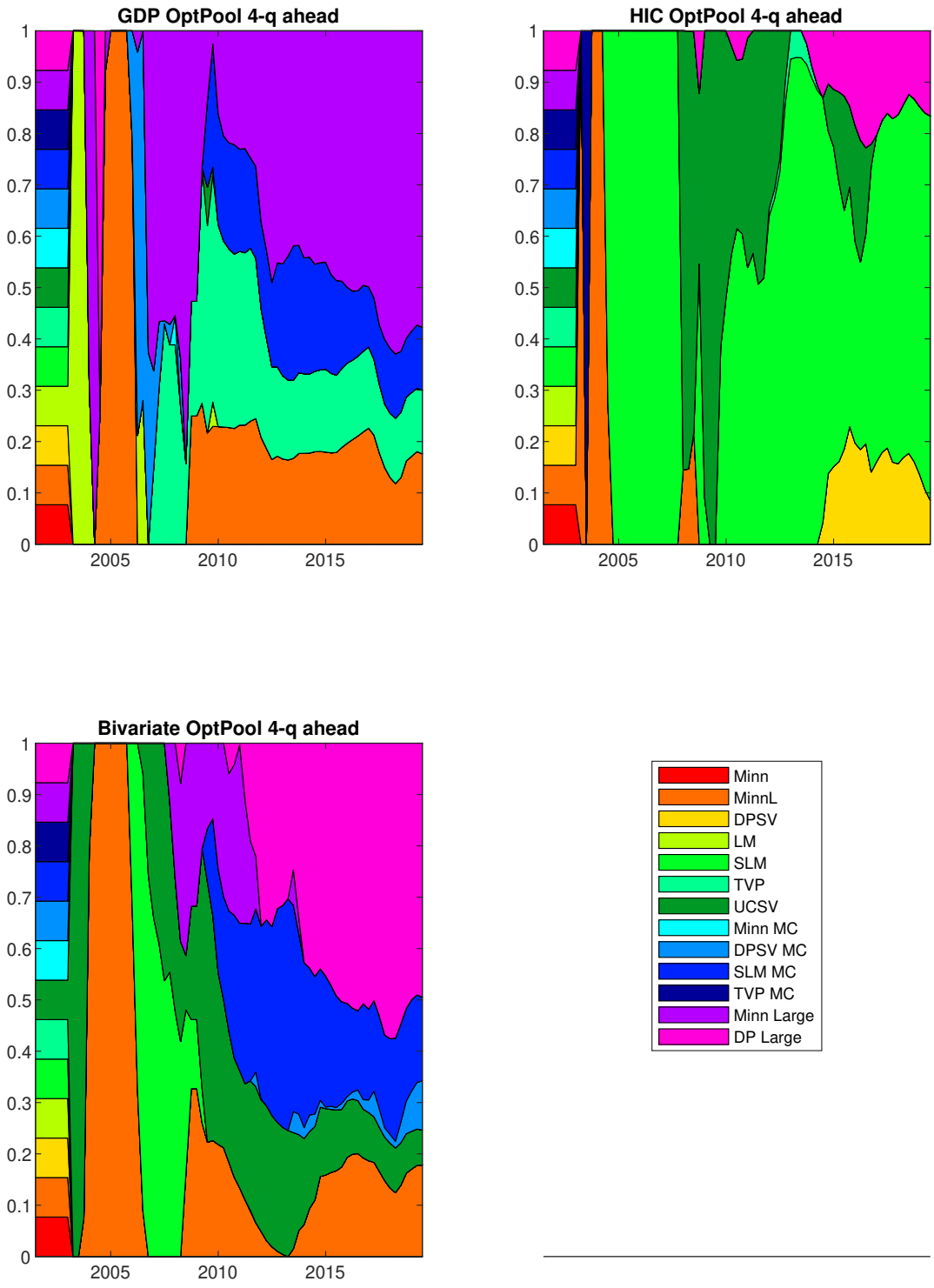


Figure 9: Weights from ex-ante mean-tilted univariate and bivariate optimal pools, one-year-ahead.

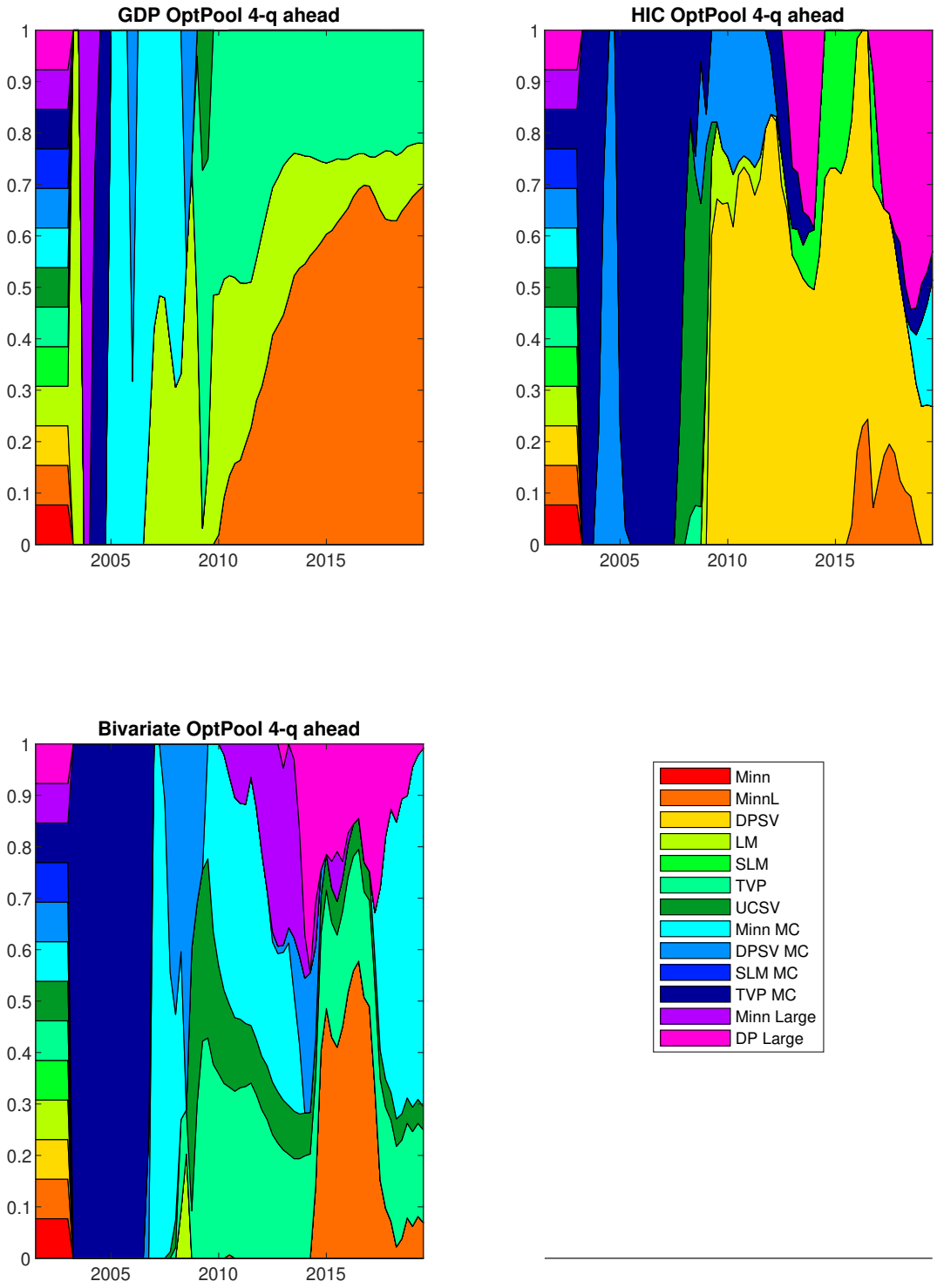
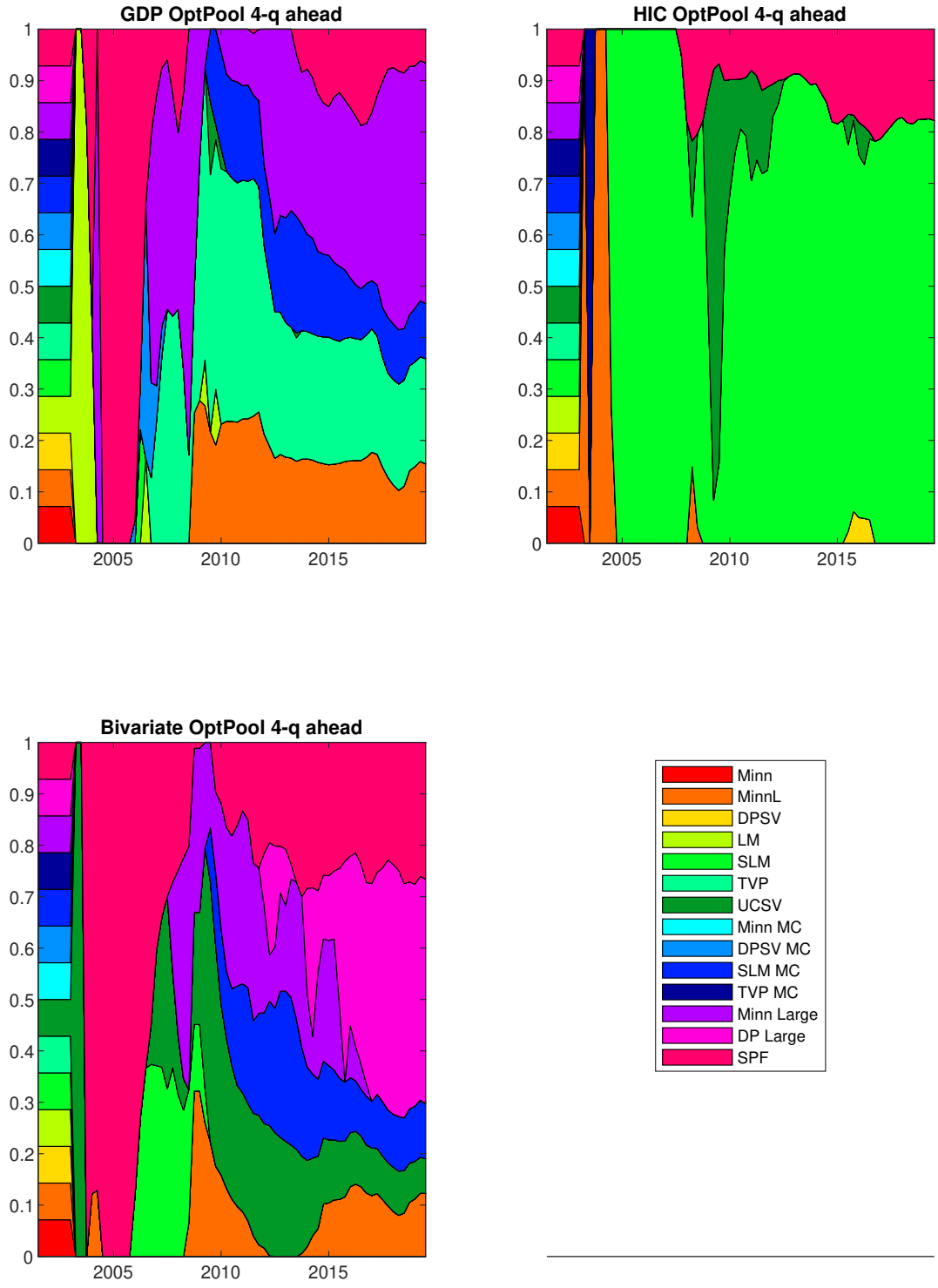


Figure 10: Weights from univariate and bivariate optimal pools including SPF, one-year-ahead.



Appendix

A Data Set

Table A.1: Data set

Variable	Small Model	Medium Model	Transformation
GDP, real	x	x	log-differences
Private consumption, real		x	log-differences
Total investment, real		x	log-differences
Exports XA, real		x	log-differences
Imports XE, real		x	log-differences
GDP deflator		x	log-differences
Total employment		x	log-differences
Short-term interest rate	x	x	levels
Long-term interest rate		x	levels
Lending rate		x	levels
Compensation per employee		x	log-differences
Headline HICP	x	x	log-differences
HICP excluding energy and food		x	log-differences
ESI		x	levels
Foreign demand		x	log-differences
Price of oil in EUR		x	log-differences
Nominal effective exchange rate		x	levels
US short-term interest rate		x	levels
US long-term interest rate		x	levels

Note: For the model specification "in levels", we use the "log-levels" instead of the "log-differences" transformation.

B Point performance of forecast combinations and SPF

Table B.1: Relative RMSFE

	Optimal Pool	SPF	Optimal Pool with SPF	Mean-tilted ex-ante	Mean-tilted ex-post	Mean and variance-tilted ex-ante	Mean and variance-tilted ex-post
Univariate							
GDP 4-q	1.069	0.972	0.999	0.989	0.955	0.953	0.959
GDP 8-q	1.281	1.044	1.008	1.092	1.028	1.050	1.046
HICP 4-q	0.689	0.964	1.013	0.975	0.998	0.973	0.976
HICP 8-q	0.823	0.894	1.001	0.898	0.919	0.906	0.904
Bivariate							
GDP 4-q	0.944	-	1.167	0.990	0.942	0.959	0.959
GDP 8-q	0.996	-	1.008	1.068	1.023	1.050	1.045
HICP 4-q	1.069	-	1.152	0.974	1.008	0.972	0.975
HICP 8-q	1.004	-	1.001	0.908	0.912	0.899	0.902

Note: The relative RMSFE is calculated as the ratio between each model RMFSEs and those of the univariate optimal pool (in absolute values in the first column of the upper panel). A number smaller than one indicates a preference for the combination over the optimal pool.

C Results of two-year-ahead forecasts

Figure C.1: Densities of two-year-ahead forecasts from combinations and SPF, real GDP growth.

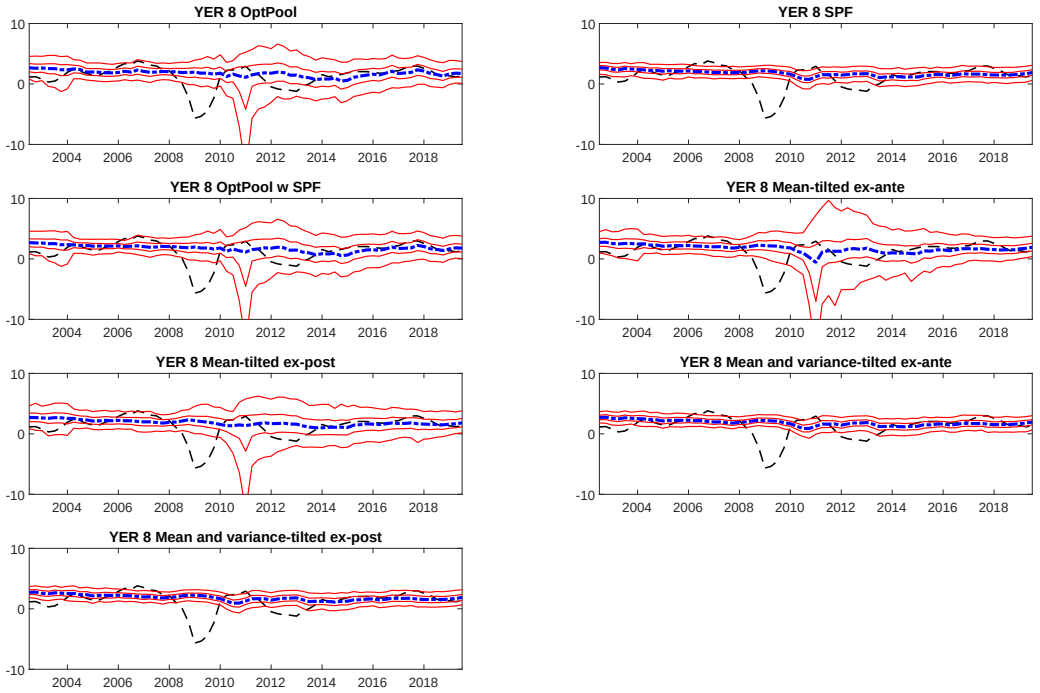


Figure C.2: PITs of two-year-ahead forecasts from combinations and SPF, real GDP growth.

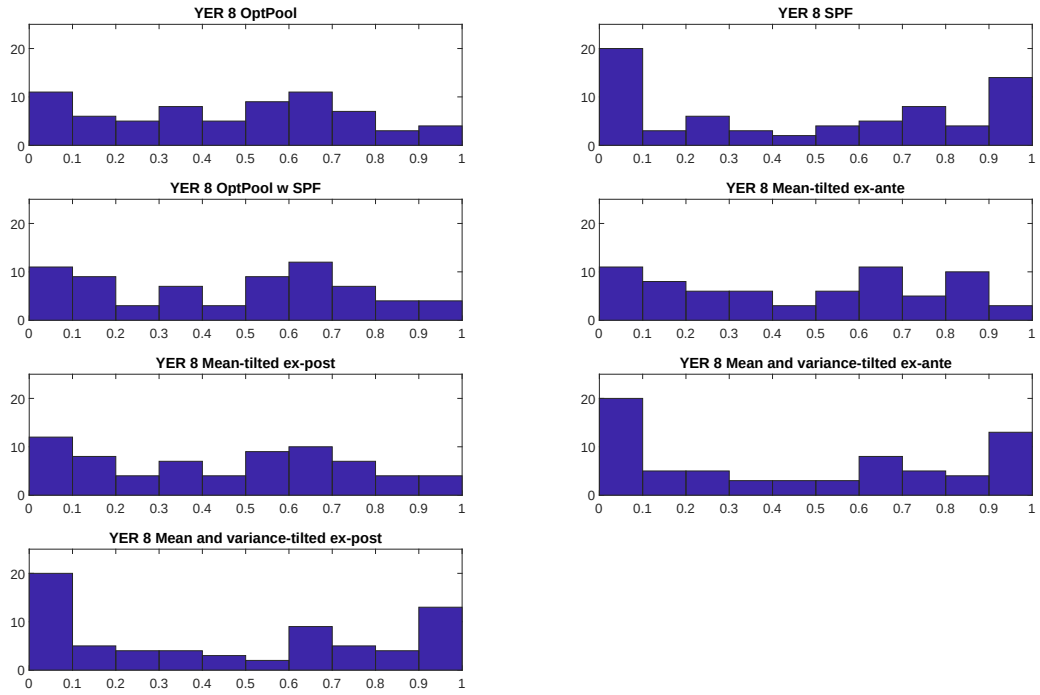


Figure C.3: Densities of two-year-ahead forecasts from combinations and SPF, HICP inflation.

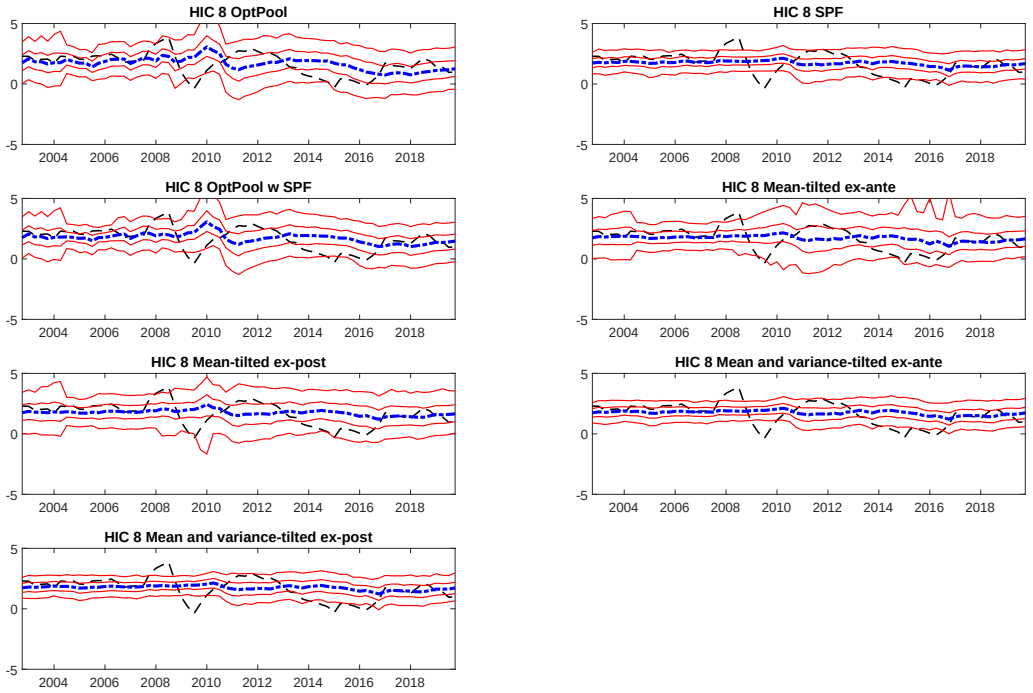


Figure C.4: PITs of two-year-ahead forecasts from combinations and SPF, HICP inflation.

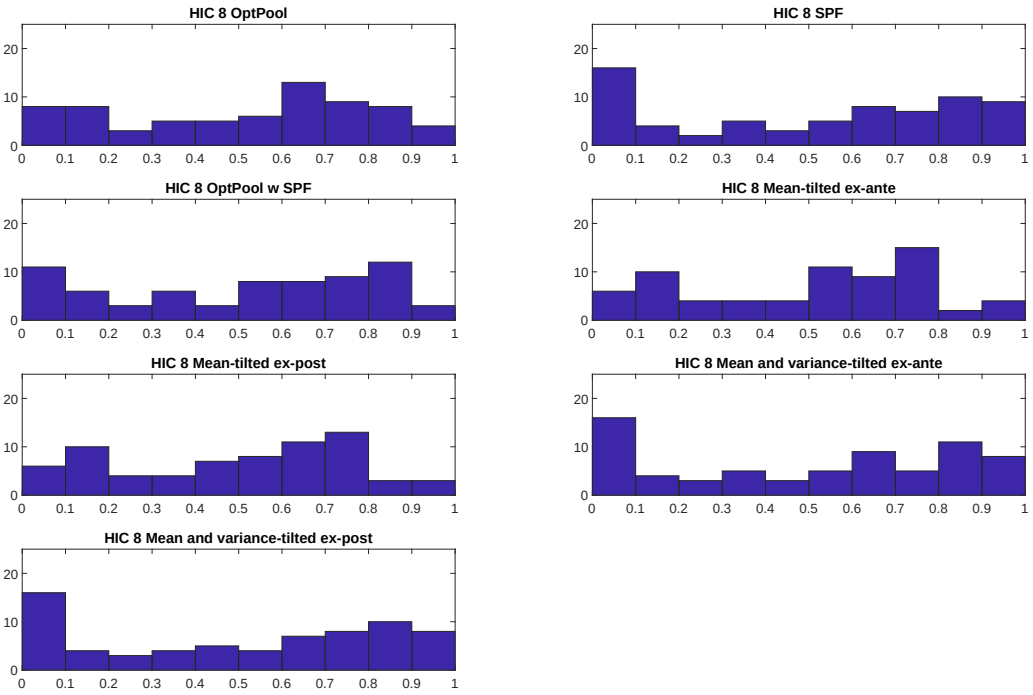


Figure C.5: Cumulative relative scores of two-year-ahead forecasts, real GDP growth (left) and HICP inflation (right).

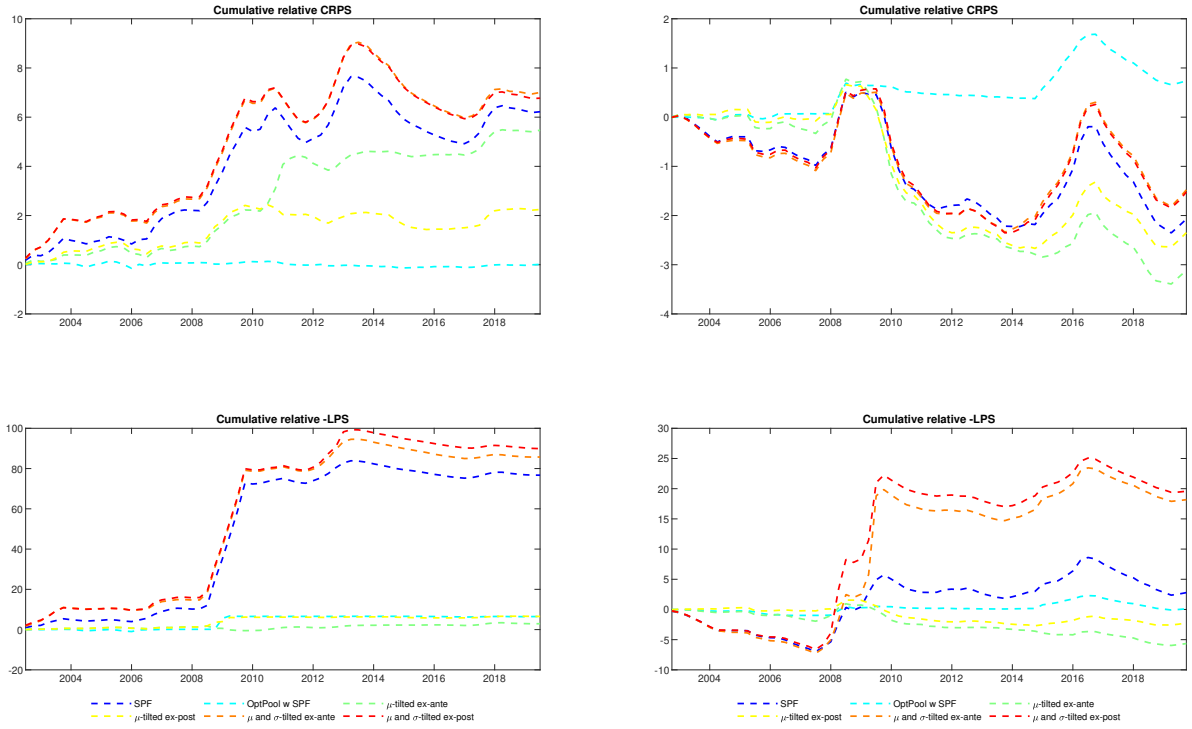


Figure C.6: Unconditional and conditional PITs for two-year-ahead real GDP growth, for two combination methods.

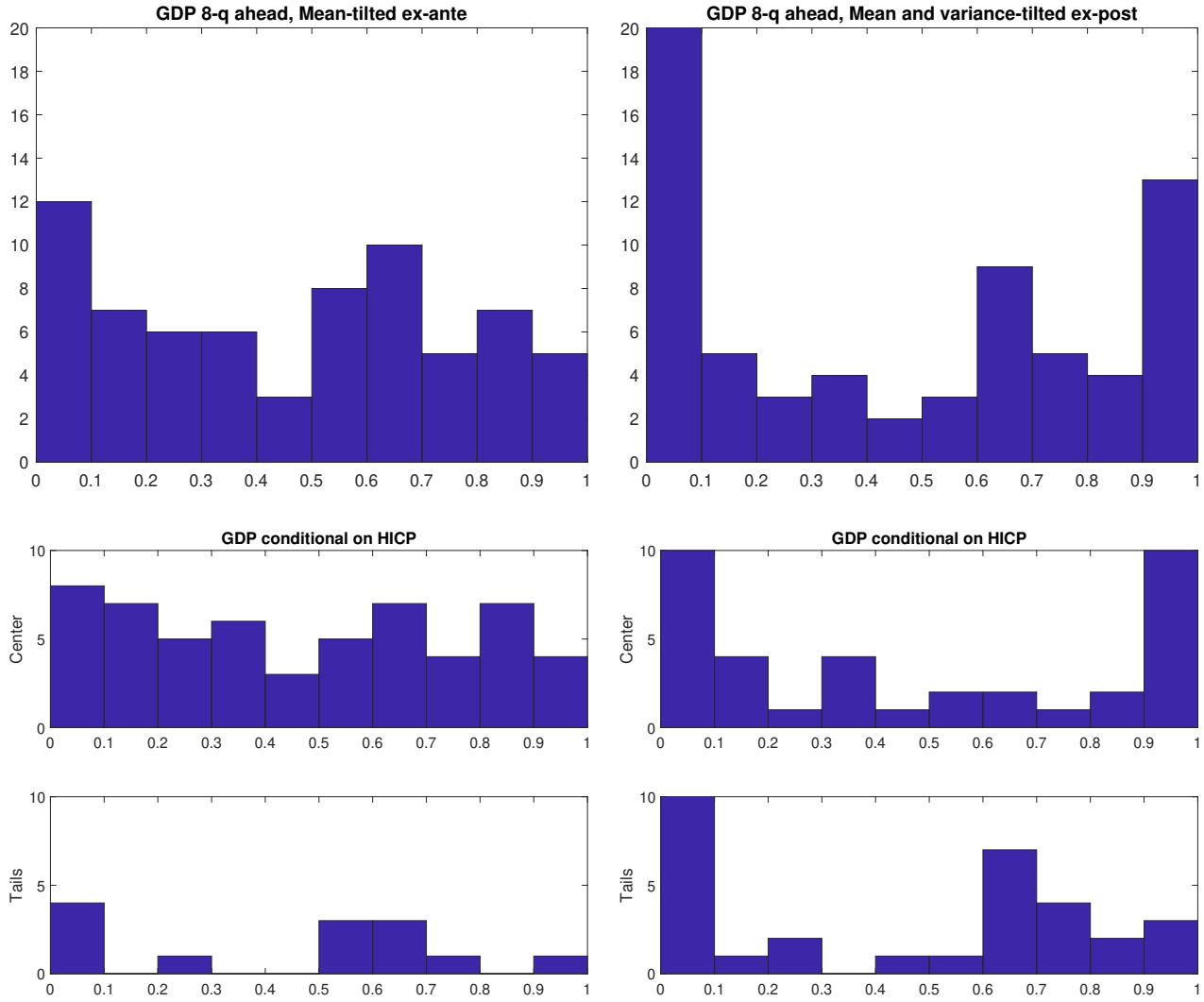


Figure C.7: Unconditional and conditional PITs for two-year-ahead HICP inflation, for two combination methods.

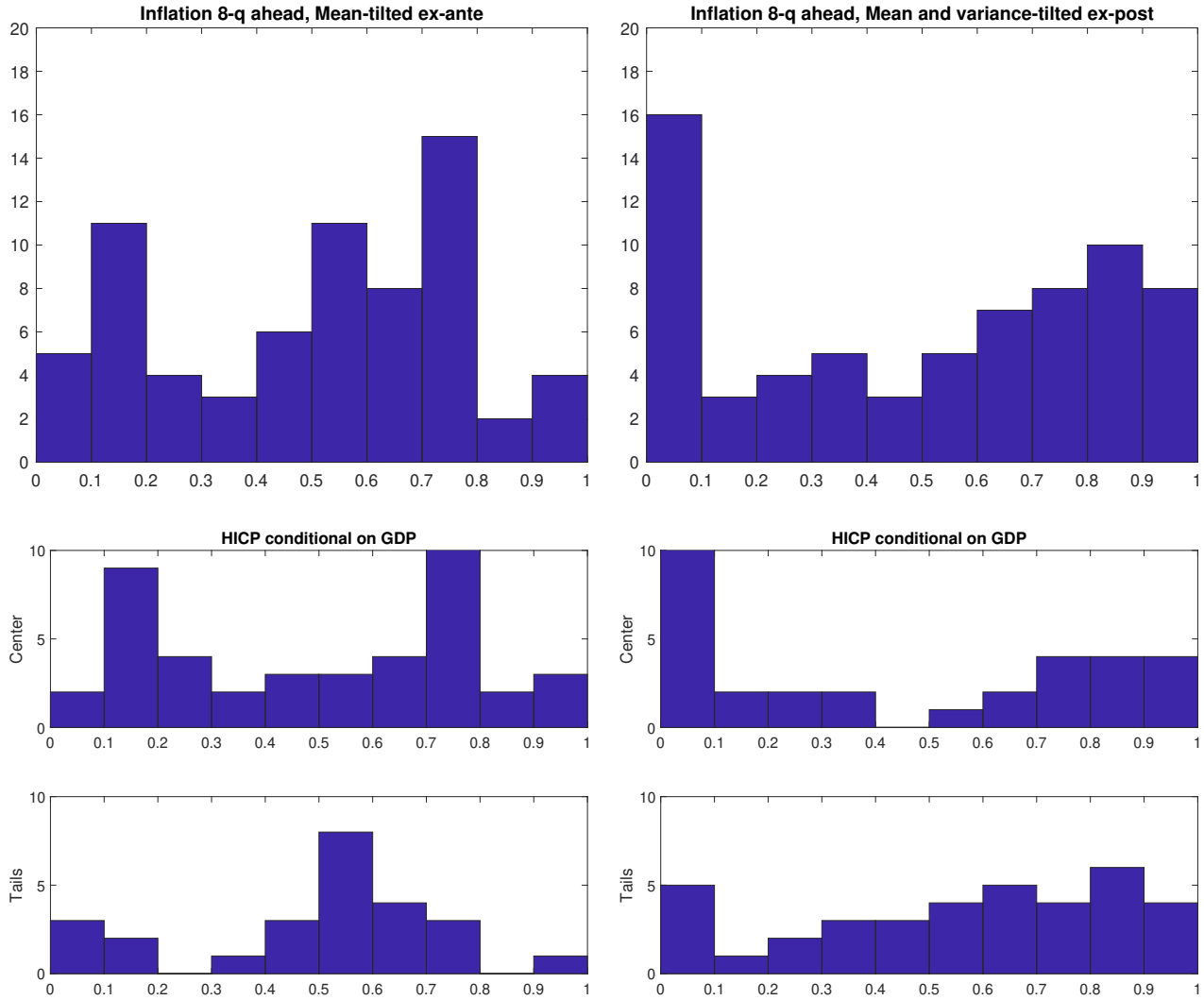


Figure C.8: Weights from univariate and bivariate optimal pools, two-year-ahead.

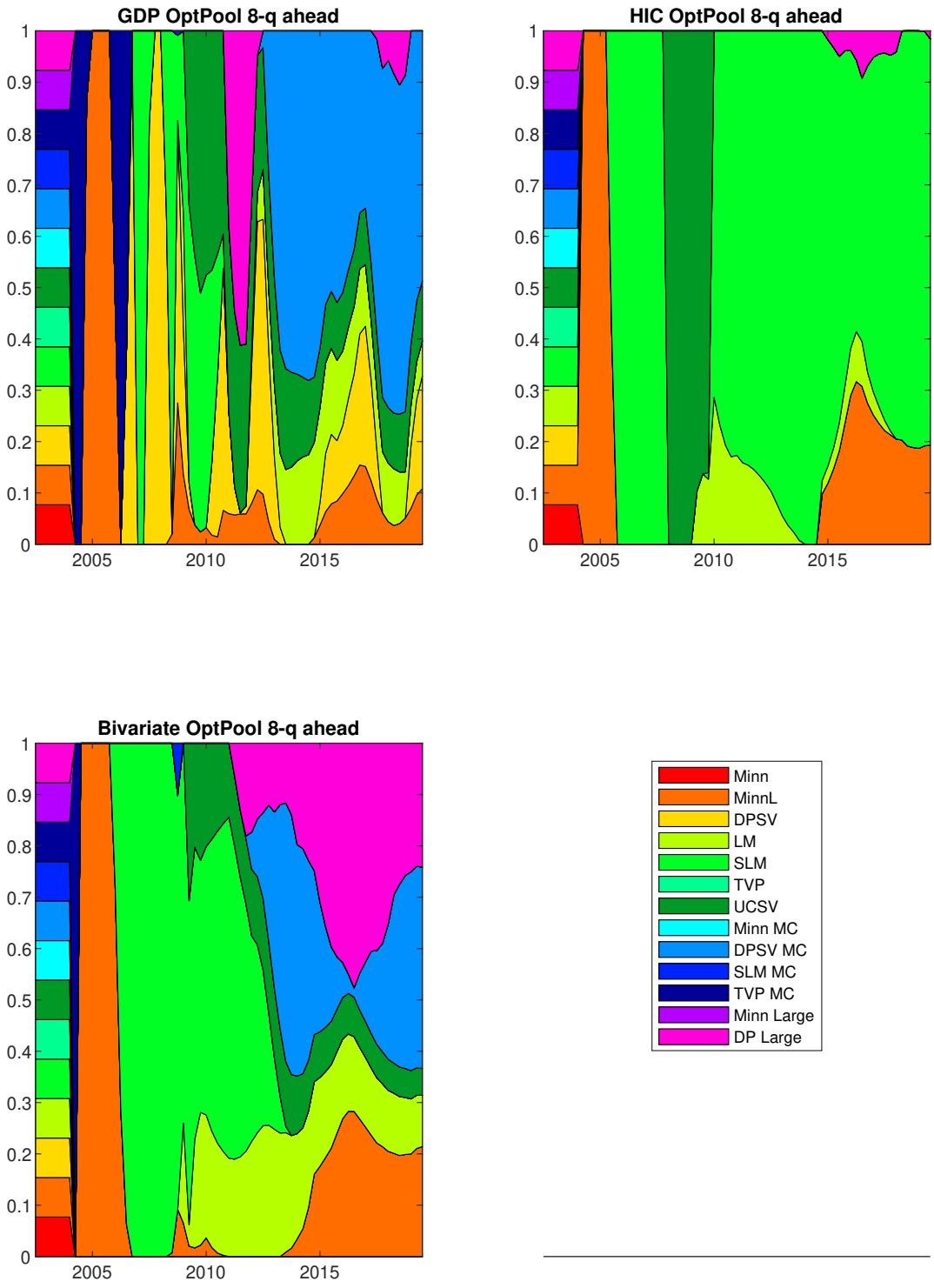


Figure C.9: Weights from ex-ante mean-tilted univariate and bivariate optimal pools, two-year-ahead.

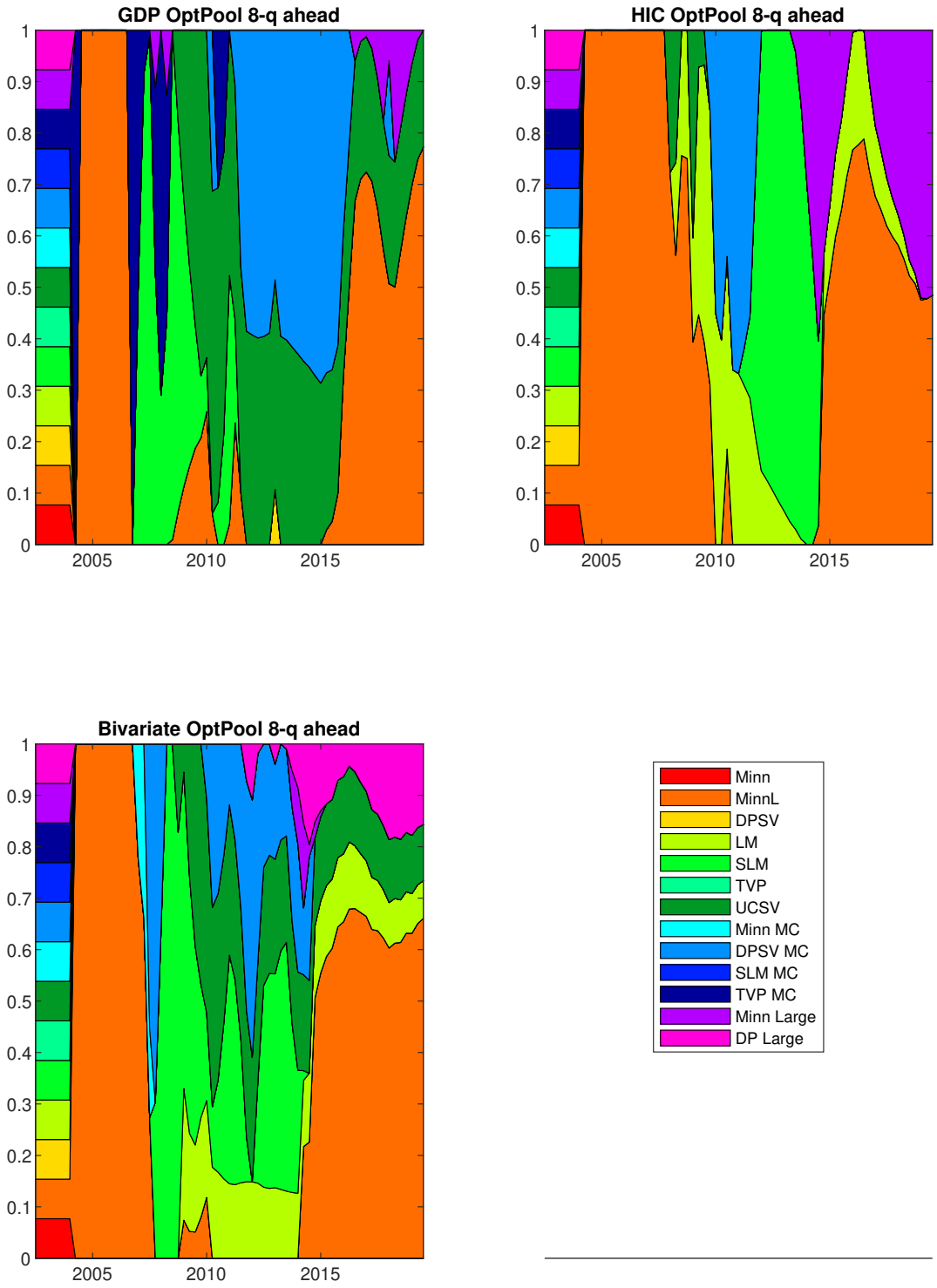
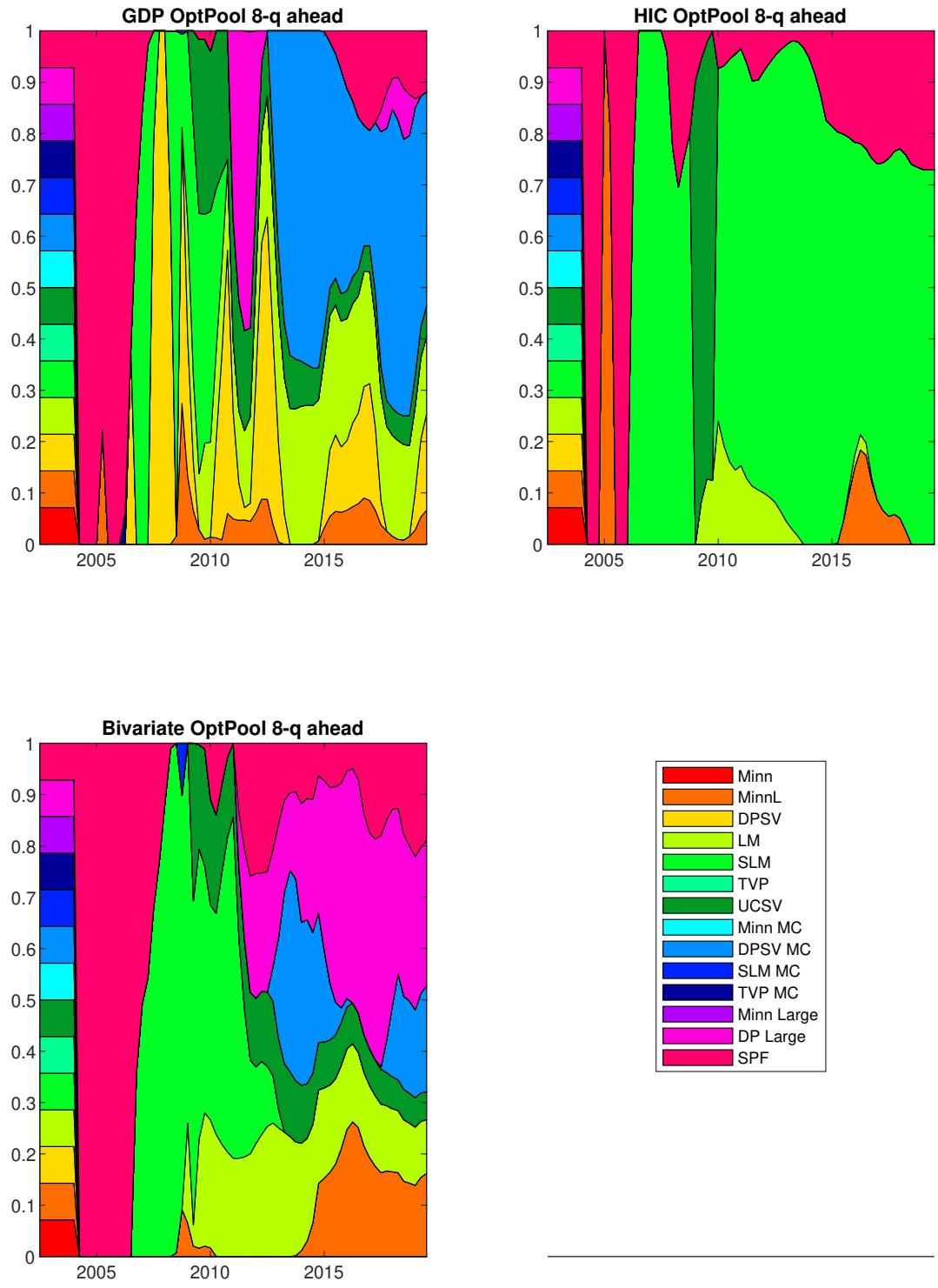


Figure C.10: Weights from univariate and bivariate optimal pools including SPF, two-year-ahead.



D Performance of individual models included in the pool

In this section, we present results for the individual BVAR models described in Section 2 of the main part. The first two Tables show accuracy scores (RMSFE, CRPS and LPS) of individual models with respect to optimal pool for the non-tilted and the mean-tilted case, for GDP and HICP, respectively. The successive Graphs D.1-D.8 show forecast distributions and PITs of each individual models for real GDP growth and HICP inflation forecasts at one- and two-year horizons.

Table D.1: Accuracy scores of individual non-tilted and mean-tilted models, relative to optimal pooling

GDP		Minn	MinnL	DPSV	LM	SLM	TVP	UCSV	Minn MC	DPSV MC	SLM MC	TVP MC	Minn Large	DP Large
4-q ahead														
Non-tilted	RMSFE	1.126	1.022	1.100	1.244	1.184	1.259	1.404	1.117	1.094	**0.996**	1.204	1.002	**0.994**
	CRPS	1.187	1.044	1.159	1.209	1.197	1.292	1.384	1.129	1.116	1.038	1.223	1.000	**0.999**
	LPS	-0.545	-0.328	-0.580	-0.182	-0.594	-0.233	-0.316	-0.586	-0.619	-0.338	-0.527	-0.223	-0.282
Mean-tilted	RMSFE	**0.958**	**0.958**	**0.959**	**0.966**	**0.949**	1.081	1.117	**0.960**	**0.951**	**0.939**	1.033	**0.955**	**0.955**
	CRPS	**0.953**	**0.933**	**0.950**	**0.950**	**0.951**	1.014	1.080	**0.937**	**0.941**	**0.950**	**0.984**	**0.937**	**0.937**
	LPS	-0.366	-0.085	-0.414	**0.004**	-0.437	**0.135**	-0.119	-0.284	-0.493	-0.232	-0.247	-0.255	-0.201
8-q ahead														
Non-tilted	RMSFE	1.039	**0.941**	1.004	1.332	1.148	1.179	1.609	1.031	**0.972**	1.027	1.156	1.059	1.058
	CRPS	1.091	**0.930**	1.056	1.235	1.125	1.215	1.478	1.034	**0.987**	1.134	1.160	1.058	1.069
	LPS	-0.852	-0.528	-0.767	-0.253	-0.684	-0.816	-0.413	-0.860	-0.774	-0.783	-0.911	-0.675	-0.647
Mean-tilted	RMSFE	1.047	1.034	1.042	1.049	1.045	1.098	1.352	1.045	1.039	1.017	1.087	1.044	1.041
	CRPS	1.064	1.042	1.058	1.073	1.060	1.111	1.308	1.056	1.054	1.151	1.094	1.053	1.059
	LPS	-0.798	-0.893	-0.836	-0.181	-0.687	-0.671	-0.311	-0.894	-0.906	-0.826	-0.975	-0.729	-0.904

Note: Ratios of scores individual model/optimal pool. For RSMFE and CRPS, a score smaller than one indicates that the individual model does better than the optimal pool. LPS is calculated as the difference between the two scores, therefore a positive value indicates a preference for the individual model over optimal pooling.

Table D.2: Accuracy scores of individual non-tilted and mean-tilted models, relative to optimal pooling

HICP		Minn	MinnL	DPSV	LM	SLM	TVP	UCSV	Minn MC	DPSV MC	SLM MC	TVP MC	Minn Large	DP Large
4-q ahead														
Non-tilted	RMSFE	1.088	1.136	1.048	1.123	1.016	1.194	1.039	1.113	1.087	1.146	1.184	1.279	1.225
	CRPS	1.072	1.132	1.030	1.132	**0.992**	1.184	1.067	1.075	1.045	1.118	1.194	1.226	1.172
	LPS	-0.102	-0.164	-0.059	-0.154	-0.015	-0.213	-0.089	-0.169	-0.112	-0.312	-0.550	-0.252	-0.185
Mean-tilted	RMSFE	**0.980**	**0.961**	**0.972**	**0.978**	**0.978**	**0.987**	**0.960**	**0.981**	**0.981**	**0.980**	1.000	**0.985**	**0.981**
	CRPS	**0.939**	**0.916**	**0.928**	**0.953**	**0.929**	1.000	**0.970**	**0.926**	**0.926**	**0.934**	**0.953**	**0.928**	**0.925**
	LPS	**0.044**	**0.085**	**0.070**	-0.008	**0.074**	-0.061	-0.016	**0.073**	**0.079**	-0.136	-0.201	**0.066**	**0.086**
8-q ahead														
Non-tilted	RMSFE	1.179	1.092	1.070	1.222	**0.983**	1.229	1.129	1.039	**0.994**	1.052	1.201	1.203	1.129
	CRPS	1.152	1.140	1.062	1.295	1.007	1.329	1.201	1.043	1.016	1.090	1.248	1.178	1.116
	LPS	-0.148	-0.210	-0.079	-0.285	-0.024	-0.345	-0.229	-0.120	-0.047	-0.135	-0.338	-0.175	-0.137
Mean-tilted	RMSFE	**0.915**	**0.878**	**0.905**	**0.917**	**0.903**	**0.984**	**0.990**	**0.907**	**0.900**	**0.910**	**0.978**	**0.906**	**0.909**
	CRPS	**0.976**	**0.906**	**0.957**	1.049	**0.942**	1.143	1.086	**0.931**	**0.931**	**0.954**	1.005	**0.935**	**0.936**
	LPS	-0.049	**0.062**	-0.015	-0.176	**0.025**	-0.265	-0.191	**0.027**	**0.045**	-0.008	-0.062	**0.049**	**0.042**

Note: Ratios of scores individual model/optimal pool. For RSMFE and CRPS, a score smaller than one indicates that the individual model does better than the optimal pool. LPS is calculated as the difference between the two scores, therefore a positive value indicates a preference for the individual model over optimal pooling.

Figure D.1: Densities of one-year-ahead forecasts from individual models, real GDP growth.

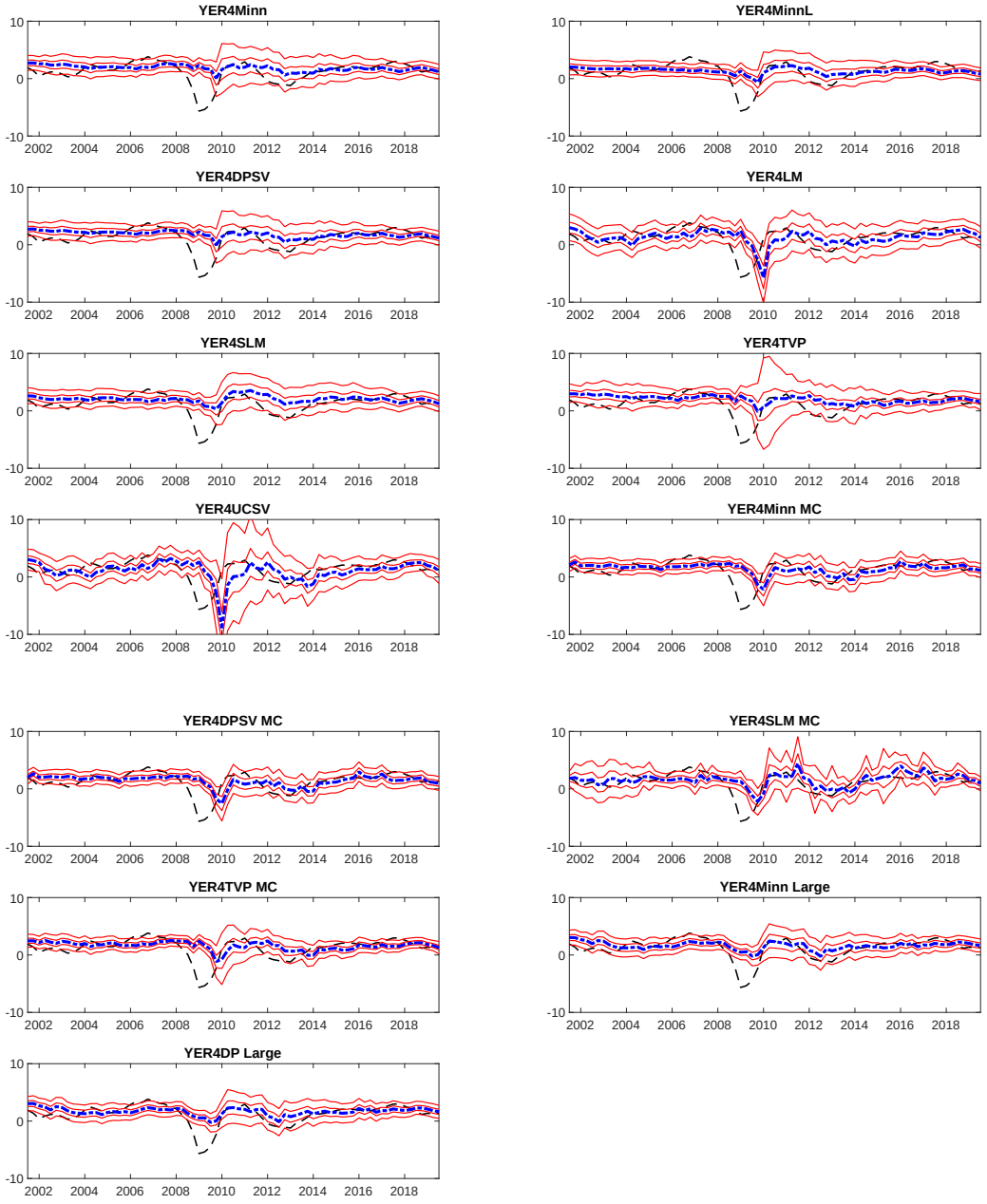


Figure D.2: PITs of one-year-ahead forecasts from individual models, real GDP growth.

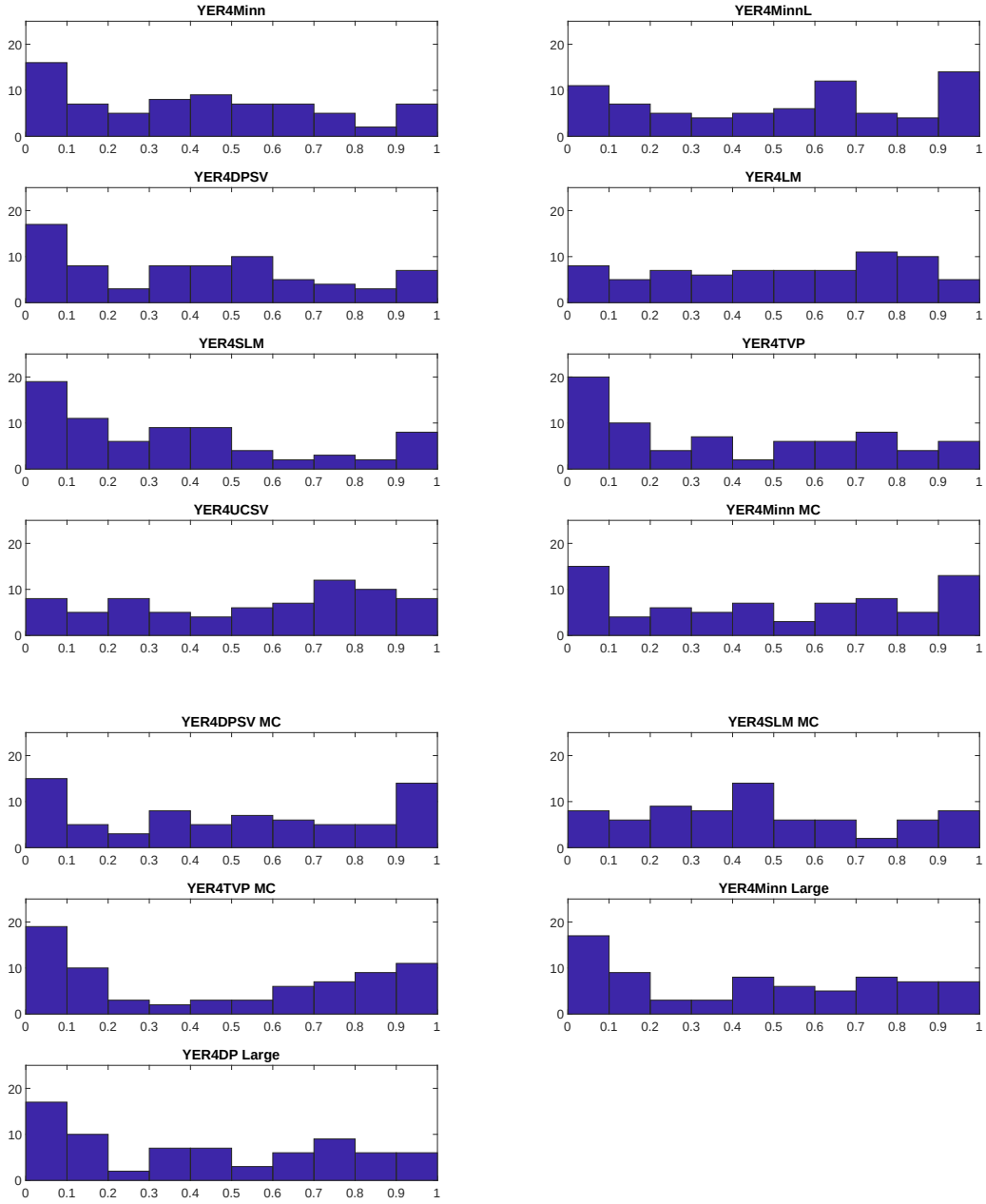


Figure D.3: Densities of one-year-ahead forecasts from individual models, HICP inflation.

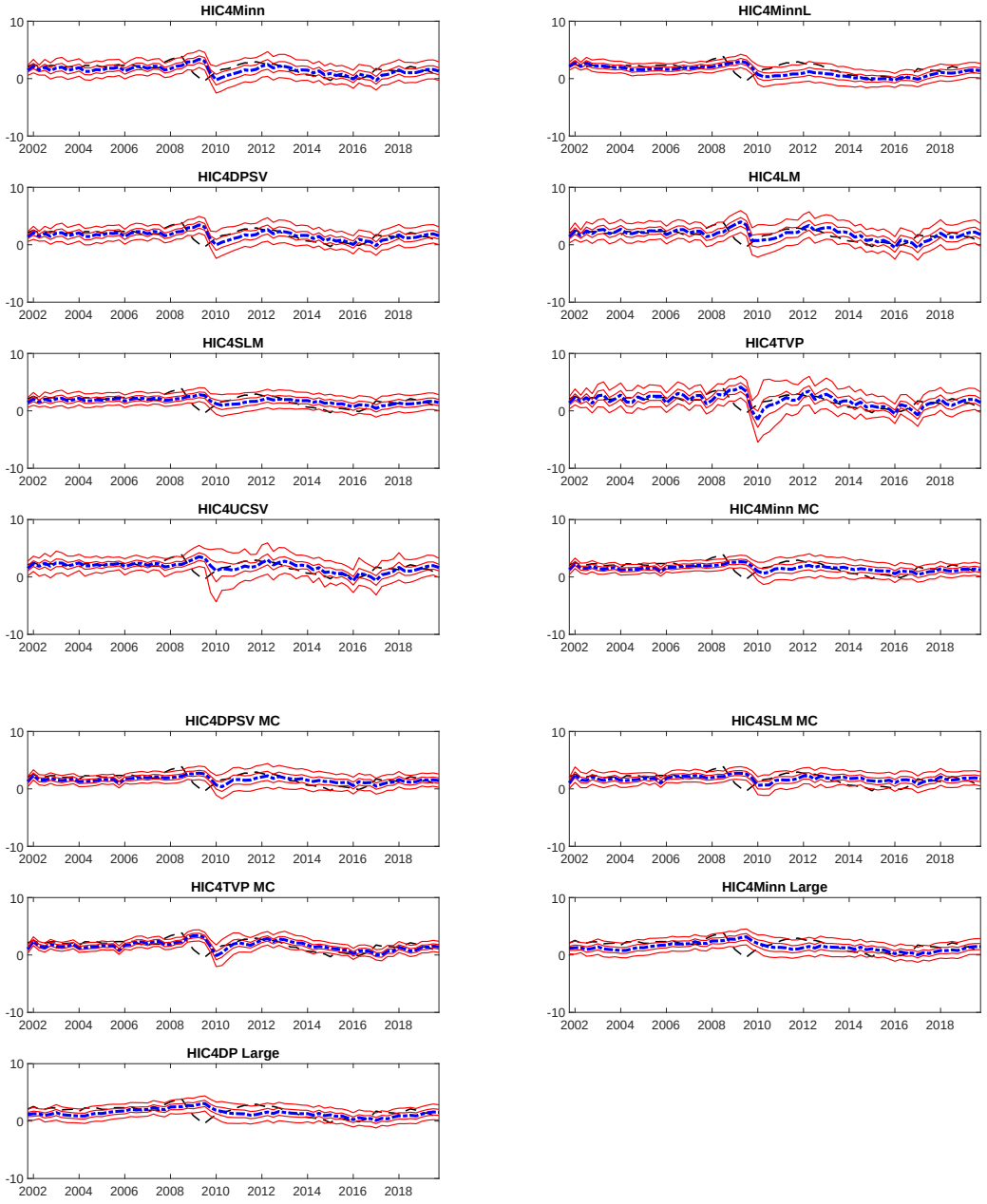


Figure D.4: PITs of one-year-ahead forecasts from individual models, HICP inflation.

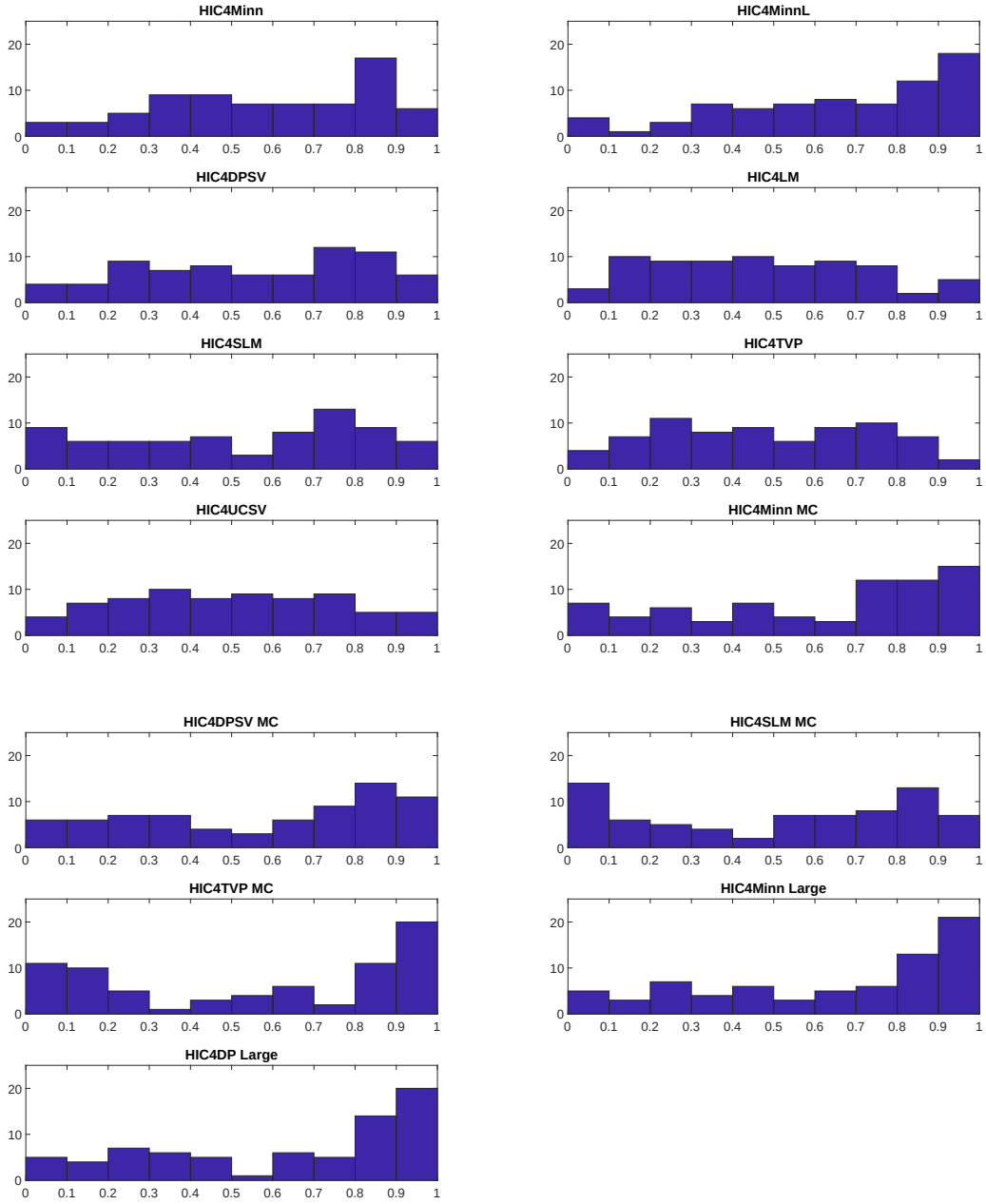


Figure D.5: Densities of two-year-ahead forecasts from individual models, real GDP growth.

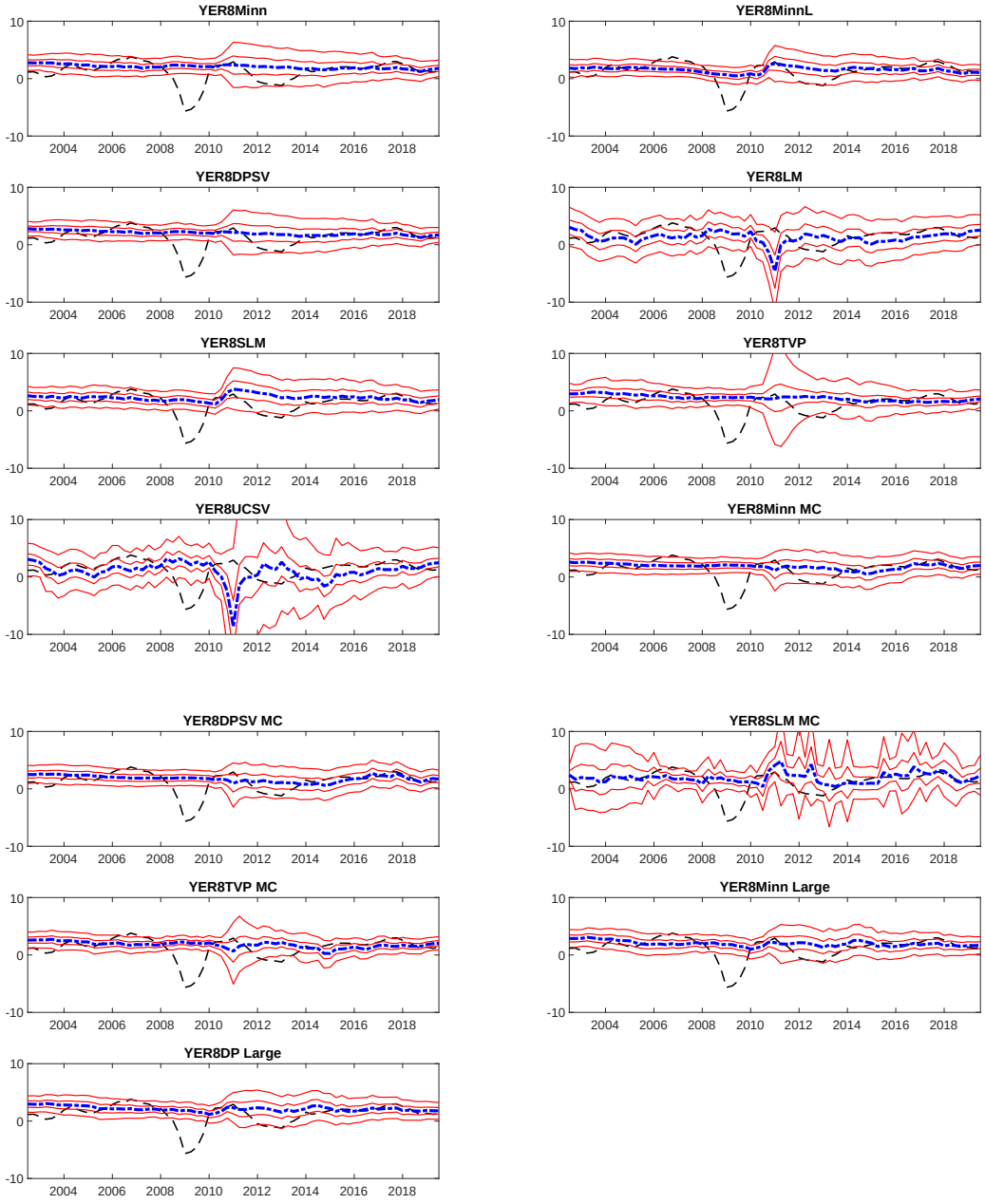


Figure D.6: PITs of two-year-ahead forecasts from individual models, real GDP growth.

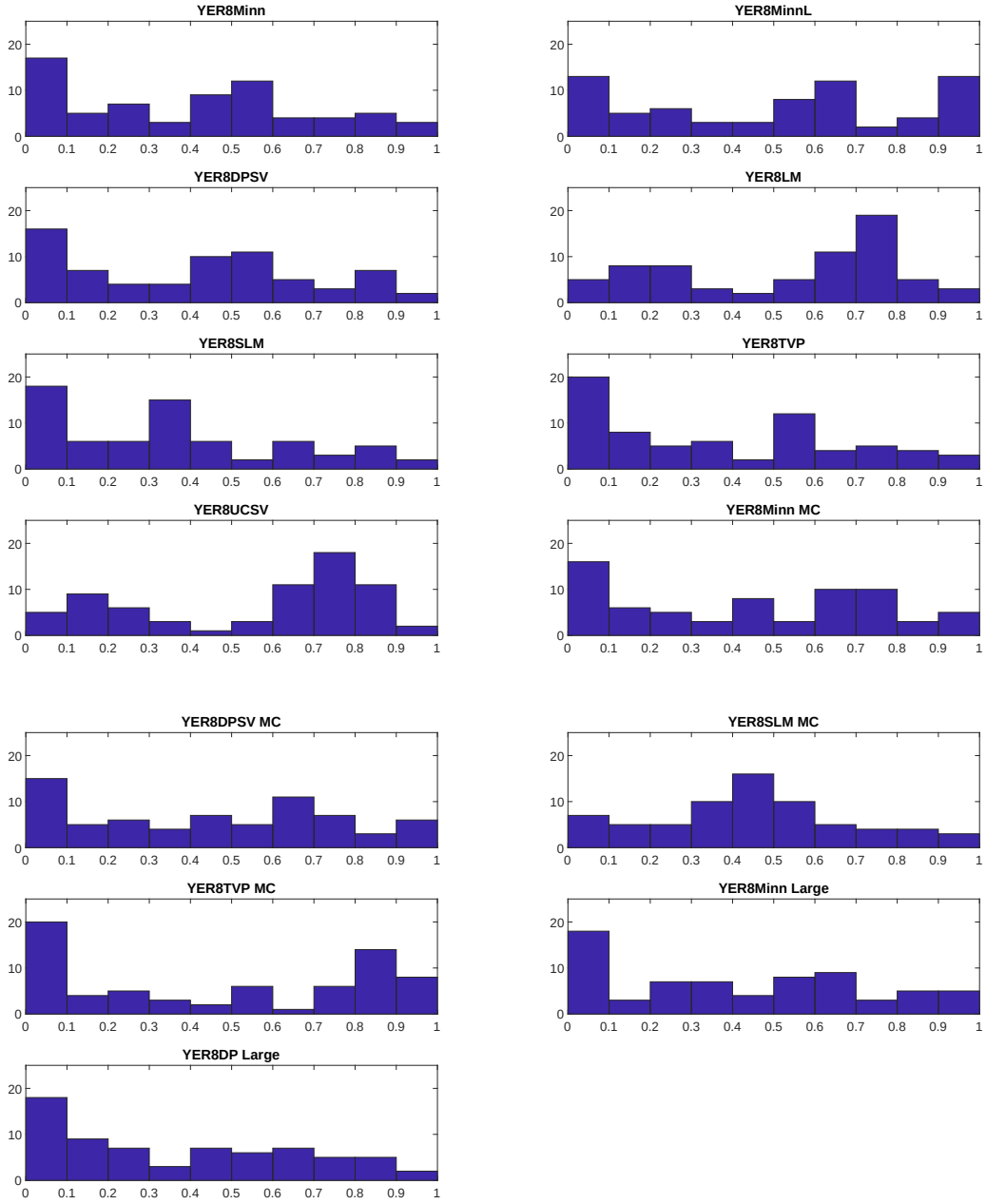


Figure D.7: Densities of two-year-ahead forecasts from individual models, HICP inflation.

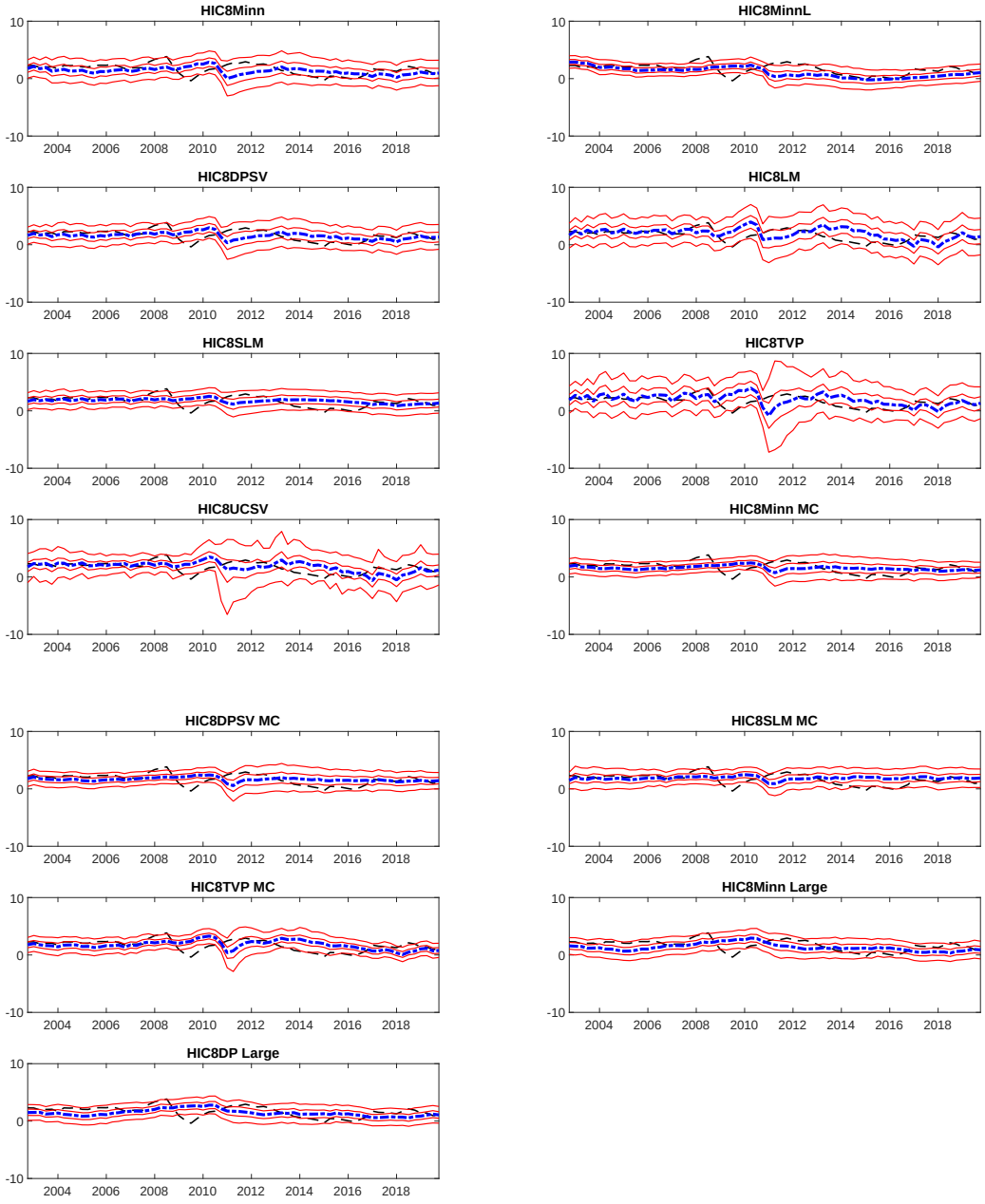


Figure D.8: PITs of two-year-ahead forecasts from individual models, HICP inflation.