Chernis, Tony; Hauzenberger, Niko; Huber, Florian; Koop, Gary; Mitchell, James

Working Paper

Predictive density combination using a tree-based synthesis function

Bank of Canada Staff Working Paper, No. 2023-61

Provided in Cooperation with:
Bank of Canada, Ottawa

Suggested Citation: Chernis, Tony; Hauzenberger, Niko; Huber, Florian; Koop, Gary; Mitchell, James (2023) : Predictive density combination using a tree-based synthesis function, Bank of Canada Staff Working Paper, No. 2023-61, Bank of Canada, Ottawa, https://doi.org/10.34989/swp-2023-61

This Version is available at:
https://hdl.handle.net/10419/297446

BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel—2023-61

Last updated: December 28, 2023

Predictive Density Combination Using a Tree-Based Synthesis Function

by Tony Chernis,[1] Niko Hauzenberger,[2, 3] Florian Huber,[3] Gary Koop[2] and James Mitchell[4]

[1] Canadian Economic Analysis Department
Bank of Canada
tchernis@bank-banque-canada.ca

[2] University of Strathclyde
United Kingdom
niko.hauzenberger@strath.ac.uk, gary.koop@strath.ac.uk

[3] University of Salzburg
Austria
florian.huber@plus.ac.at

[4] Federal Reserve Bank of Cleveland
United States
James.Mitchell@clev.frb.org

DOI: https://doi.org/10.34989/swp-2023-61 | **ISSN** 1701-9397

Acknowledgements

The views expressed are solely those of the authors and do not necessarily reflect the views of the Bank of Canada, the Federal Reserve Bank of Cleveland, or the Federal Reserve System. We thank Mike West and conference and seminar participants at Notre Dame and Örebro, including Christiane Baumeister, Drew Creal, Luca Rossini, and Mattias Villani for helpful comments. Niko Hauzenberger gratefully acknowledges financial support from the Austrian Science Fund (FWF, ZK-35) and the Austrian Central Bank (Anniversary Fund, project no. 18763). Florian Huber gratefully acknowledges financial support from the Austrian Science Fund (FWF, ZK-35).

Abstract

Bayesian predictive synthesis (BPS) is a method of combining predictive distributions based on agent opinion analysis theory, which encompasses many common approaches to combining density forecasts. The key ingredient in BPS is a synthesis function. This is typically specified parametrically as a dynamic linear regression. In this paper, we develop a nonparametric treatment of the synthesis function using regression trees. We show the advantages of our tree-based approach in two macroeconomic forecasting applications. The first uses density forecasts for GDP growth from the euro area's Survey of Professional Forecasters. The second combines density forecasts of US inflation produced by many regression models involving different predictors. Both applications demonstrate the benefits—in terms of improved forecast accuracy and interpretability—of modeling the synthesis function nonparametrically.

Topic: Econometric and Statistical Methods
JEL codes: C11, C32, C53

Résumé

La synthèse de prévisions bayésienne est une méthode servant à combiner des distributions prédictives qui est basée sur la théorie d'analyse des avis d'agents et qui englobe de nombreuses approches courantes pour la combinaison de prévisions par densité de probabilités. L'élément clé de cette méthode consiste en une fonction de synthèse qui est généralement définie de façon paramétrique sous forme d'une régression linéaire dynamique. Dans cette étude, nous élaborons un mode de traitement non paramétrique de la fonction de synthèse fondé sur des arbres de régression. Nous montrons les avantages que notre mode de traitement présente dans le cadre de deux exercices de prévision macroéconomique. Dans le premier exercice, nous utilisons des prévisions par densité de probabilités de la croissance du produit intérieur brut issues de l'enquête menée auprès des prévisionnistes professionnels de la zone euro. Dans le second, nous combinons des prévisions par densité de probabilités de l'inflation aux États-Unis, qui sont produites par de nombreux modèles de régression comportant différents prédicteurs. Ces deux exercices démontrent les avantages que procure la modélisation non paramétrique de la fonction de synthèse, sur le plan de l'amélioration de l'exactitude et de l'interprétabilité des prévisions.

Sujet : Méthodes économétriques et statistiques
Codes JEL : C11, C32, C53

1 Introduction

It is commonplace when forecasting macroeconomic variables, such as output growth or inflation, to consider a large number of competing predictive densities. These density forecasts might come from different reduced-form or structural models and/or be subjective and come from surveys. How to combine these densities is an open question being addressed by a growing literature.[1] The literature concludes that combined density forecasts tend to be more accurate and more robust than single-model approaches that ignore model uncertainty; for a review, see Aastveit et al. (2019). One issue is that traditional forecast combination techniques are often linear and do not exploit information besides the forecasts and the target variable. Contrast this with a policymaker who combines forecasts nonlinearly and uses external information, such as on the current state of the economy or financial conditions, to help determine how much weight to attach to the different forecasts. We propose a novel technique that mimics this practice. Our approach combines density forecasts nonparametrically while allowing the combination weights to be determined by information that may be external to the forecasting models.

Key to our approach is Bayesian predictive synthesis (BPS). It has emerged, as extended into a time-series context by McAlinn and West (2019), as a general method of density forecast combination with a strong theoretical basis. BPS draws on an earlier Bayesian literature on agent or expert opinion analysis (West and Crosse, 1992) and provides a formal and theoretically justified method for pooling densities. It can be shown to nest many previous approaches (see, for example, Section 2.2 of McAlinn and West, 2019) and has been used successfully in various applications in economics, such as McAlinn et al. (2020), Chernis (2023), and Aastveit et al. (2023). In this paper we develop density forecast combination strategies within the BPS framework.

In existing implementations of BPS, the so-called synthesis function, which determines the weight attached to each density, needs to be specified parametrically. Common choices, as made in the aforementioned papers, are to assume that the synthesis function takes the form of a dynamic linear regression, with parameters that are allowed to change over time typically as random walk processes. This specification of the synthesis function thus allows the weights

[1] See, among many others, Mitchell and Hall (2005); Wallis (2005); Hall and Mitchell (2007); Geweke and Amisano (2011); Koop and Korobilis (2012); Billio et al. (2013); Aastveit et al. (2014); Conflitti et al. (2015); Chernis and Webley (2022); Knotek and Zaman (2023); Aastveit et al. (2023); Čapek et al. (2023); Diebold et al. (2023).

on competing density forecasts to evolve over time as linear Gaussian random walks. But such an assumption may or may not be valid. Misspecification occurs if the weights depend on other factors or if they follow a different law of motion than a random walk.

These considerations motivate the present paper. BPS has theoretically rigorous foundations, but the manner in which it has been implemented in practice risks misspecification due to the adoption of particular and untested parametric assumptions. We therefore propose to use flexible nonparametric techniques to specify the synthesis function. Specifically, we use regression trees. In conventional (single-model) forecasting applications, tree-based models of the conditional mean have proven highly successful (see, for example, Clark et al., 2023; Huber and Rossini, 2022; Huber et al., 2023). A small number of other papers have used nonparametric techniques to combine predictive densities (for example, Jin et al., 2022; Bassetti et al., 2018, 2023). However, unlike our proposed method, these other papers neither use regression trees nor fit explicitly within the formal BPS framework.

While regression trees have become a popular way to estimate nonparametric regressions, here we propose to use them differently. Similarly to Coulombe (2020), Deshpande et al. (2020), and Hauzenberger et al. (2023), who provide a nonparametric treatment to the parameters rather than the variables in single-equation and VAR models, we model the coefficients in the synthesis function with regression trees (RT). Accordingly, we label our version of BPS "BPS-RT." The synthesis function remains linear in the parameters, which, as we will demonstrate, aids in interpretation. Use of regression-tree methods requires the choice of covariates, which we call "weight modifiers." These weight modifiers help determine the weights attached to the competing density forecasts. Conventional BPS does not make use of weight modifiers, given that the weights are typically assumed to follow random walks. Thus, in popular implementations of BPS, any relevant information in the form of additional covariates is neglected.[2] But decision makers, when combining competing density forecasts, may wish to condition their forecasts on such "outside" information. For example, they may wish to let the weights on the different forecasting models vary with the state of the economy or vary as a function of the features of each forecast density. Our tree-based specification for the synthesis function is able to condition on

[2]Notable recent exceptions are Oelrich et al. (2023), who, following Li et al. (2023), let the weights in linear density forecast combinations depend on (potentially time-varying) exogenous variables. As Oelrich et al. (2023) explain, such linear pools are one specific instance of BPS. Letting the combination weights in linear pools change over time according to these "pooling variables," as in the more general (nonlinear) BPS framework that we consider, can offer more flexibility than assuming that the combination weights follow an assumed autoregressive process; cf. Del Negro et al. (2016).

both "global" (that is, information not associated with a particular forecaster) and "local" (that is, information associated with a given forecaster) variables when determining the weights. In our tree-based synthesis function, the weights on each density forecast are dynamically determined via a sequence of decision rules. BPS-RT allows the decision maker to combine predictive densities in a highly flexible way and to distill optimally all relevant information contained in the predictive densities and weight modifiers. The fact that the synthesis function remains conditionally linear in the parameters helps the decision maker interpret the combined density and better understand the role each individual density is playing in the combination. We will show how BPS-RT can be used to understand the role of model incompleteness, agent clustering, and the time-varying importance of the different weight modifiers.

The next section of the paper introduces and motivates BPS in theory and then discusses how it has been implemented in the existing literature. It then proposes our generalization, BPS-RT, and explores its properties. Section 3 demonstrates the utility of BPS-RT by undertaking two forecasting applications. The first application takes the individual forecaster density forecasts from the European Central Bank Survey of Professional Forecasters (ECB SPF) and combines them. The second application forecasts US inflation using a commonly used large set of indicators. The predictive densities that are synthesized are produced by regression models using the different indicators. We find that BPS-RT produces well-calibrated and accurate forecasts. Notably, we find that single-tree models perform best, in contrast to standard recommendations when using regression trees. This suggests that a relatively parsimonious weight scheme with few changes in weights is supported by the data. The superior performance of BPS-RT stems from its better ability to explain periods of volatility, such as the global financial crisis that affected euro area GDP growth and the post-COVID inflation period in the US. Zooming in on the best-performing BPS-RT specification in the US inflation application, we show how the combination forecasts from BPS-RT can be interpreted. BPS-RT can be used to understand the role of model incompleteness, agent (forecast) clustering, and the time-varying importance of the different weight modifiers. We find little model set incompleteness during the post-COVID inflation period, suggesting that BPS-RT's success comes from its ability to successfully forecast inflation using the underlying models with changes in the combination weights driven by a time trend. This contrasts with the earlier period of lower inflation, when business cycle indicators are shown to be more important. Section 4 concludes. Online Appendix A provides details

on Bayesian inference of BPS-RT, and Appendix B provides additional empirical results, as referenced in the main paper.

2 Bayesian Predictive Synthesis with Regression Trees

In Section 2.1, we provide some background on BPS, distinguishing between BPS in theory and its use in practice in extant empirical applications. Then, in Section 2.2, we explain how regression trees can be used to provide a more flexible way of operationalizing BPS.

2.1 Bayesian Predictive Synthesis

2.1.1 BPS in Theory

BPS is a foundational theoretically coherent Bayesian method for combining predictive densities.[3] The theory of BPS provides a pooled predictive distribution for the variable being forecast (say, GDP growth) given a set of individual density forecasts. Operationally, this pooled predictive distribution is produced using Markov chain Monte Carlo (MCMC) methods involving two steps. In the first step, draws are taken from the individual predictive densities for GDP growth. These draws are then, in effect, treated in a second step as explanatory variables in a time-series model where the dependent variable is the outcomes for GDP growth. This time-series model amounts to the synthesis function. Standard choices for this function are typically based on linearity, either simply a constant linear relationship or a dynamic relationship where the linear coefficients evolve over time according to a random walk. As pointed out by Aastveit et al. (2023), this means that BPS can be thought of as a multivariate regression relating the target variable (GDP growth) to the forecasts for GDP growth, which are treated as generated regressors. We make use of this generated regressor interpretation below.

More formally, at time t a decision maker $\mathcal{D}$ is confronted with h-step-ahead forecast densities for variable $y_{\tau+h}$ produced by J different agents, experts, or models, where τ ranges from 1 to t. At each forecast origin, τ, we label these predictive densities $\{\pi_{j\tau}(y_{\tau+h})\}_{j=1}^{J}$. These densities, available from time 1 through t, form the information set $\mathcal{H}_t$ of $\mathcal{D}$ and can, in principle, be of any distributional form. $\mathcal{D}$ then forms an incomplete joint prior $p(y_{t+h}, \mathcal{H}_t) = p(y_{t+h}) \times \mathbb{E}\left(\prod_j \pi_{jt}(x_{jt+h|t})\right)$ with $\boldsymbol{x}_{t+h|t} = (x_{1t+h|t}, \ldots x_{Jt+h|t})'$ denoting latent agent states

[3] For a general description of BPS, see McAlinn and West (2019); specific implementation details related to our applications are discussed below and in Appendix A.

(that is, draws from the agent-specific predictive densities). These agents' forecasts target $t+h$ but, under the prior, are made using information through time t. The prior is incomplete, in the sense that $\mathcal{D}$ only forms an expectation of the product of the agent densities. Agent opinion analysis theory (West, 1992; West and Crosse, 1992), extended to a time-series context by McAlinn and West (2019), shows that the posterior conditional density for y_{t+h} under this incomplete prior takes the form

$$p(y_{t+h}|\boldsymbol{\Psi}_{t+h}, \mathcal{H}_t) = \int \alpha(y_{t+h}|\boldsymbol{x}_{t+h|t}, \boldsymbol{\Psi}_{t+h}) \prod_{j=1}^{J} \pi_{jt}(x_{jt+h|t}) dx_{jt+h|t}, \tag{1}$$

where $\alpha(y_{t+h}|\boldsymbol{x}_{t+h|t}, \boldsymbol{\Psi}_{t+h})$ denotes the synthesis function that reflects how $\mathcal{D}$ combines her prior information with the set of expert-based forecasts; $\boldsymbol{\Psi}_{t+h}$ denotes a matrix of parameters and latent states that control the properties of the synthesis function, $\alpha(.)$.

2.1.2 BPS in Practice

Theory offers no guide as to the specific choice of the synthesis function, $\alpha(y_{t+h}|\boldsymbol{x}_{t+h|t}, \boldsymbol{\Psi}_{t+h})$. But a common choice in empirical applications, used, for example, in McAlinn and West (2019), McAlinn et al. (2020), and Aastveit et al. (2023), is to assume a dynamic linear regression model treating the draws from the J competing densities as generated regressors, $\boldsymbol{x}_{t+h|t}$. Our synthesis functions will have a dynamic regression form, but we will use a non-centered parameterization (see Frühwirth-Schnatter and Wagner, 2010):

$$\alpha(y_{t+h}|\boldsymbol{x}_{t+h|t}, \boldsymbol{\Psi}_{t+h}) = \mathcal{N}\left(y_{t+h}|c_{t+h} + \sum_{j=1}^{J}(\gamma_j + \beta_{jt+h})x_{jt+h|t}, \sigma^2_{t+h}\right), \tag{2}$$

where c_{t+h} is a time-varying intercept assumed to follow a random walk, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)'$ are time-invariant weights, and $\boldsymbol{\beta}_{t+h} = (\beta_{1t+h}, \ldots, \beta_{Jt+h})'$ denotes the time-varying combination weights. As discussed above, a common choice in the literature is to assume that the weights, β_{jt+h}, evolve as a random walk (RW) with innovation covariance matrix $\boldsymbol{V}$, leading to a version of BPS that we label "BPS-RW." When implementing BPS-RW in our empirical applications below, we make standard choices for the prior and MCMC method. In particular, they are similar to those used in Hauzenberger et al. (2022). The only difference is that we use the hyperparameter-free horseshoe prior instead of the normal-Gamma prior, so as to have a prior

that is comparable to the one used with our regression-tree model. Accordingly, we do not provide additional details here on drawing the time-varying weights for BPS-RW; see Hauzenberger et al. (2022) for details.

In all of our implementations of BPS, including BPS-RW, we consider two versions: one that assumes stochastic volatility (SV) and another that is homoskedastic. In the SV case, the error variance, σ^2_{t+h}, changes over time. We assume that the log-volatilities $\varsigma_{t+h} := \log \sigma^2_{t+h}$ evolve according to an AR(1) model with autoregressive coefficient ρ_ς, unconditional mean μ_ς, initial value ς_0, and error variance σ^2_ς. The prior choices for these parameters are given in Appendix A.1. Homoskedastic cases are obtained setting σ^2_ς to zero. Below, for notational ease, we do not explicitly note those parameters relating to SV in the conditioning arguments.

All of our implementations of BPS also include a time-varying intercept, c_{t+h}, which is assumed to follow a random walk. As discussed below, c_{t+h} is included to allow for model incompleteness. Further econometric details are provided in Appendix A.

With these notational conventions established, $\boldsymbol{\Psi}_{t+h} = \left(\boldsymbol{\gamma}, \{c_\tau, \boldsymbol{\beta}_\tau, \sigma_\tau\}_{\tau=0}^{t+h}, \boldsymbol{\theta}\right)$, where $\boldsymbol{\theta}$ will be method-specific parameters that define the law of motion of latent states or appear in the hierarchical priors (such as $\boldsymbol{V}$ in the case of BPS-RW).

The synthesis function, $\alpha(y_{t+h}|\boldsymbol{x}_{t+h|t}, \boldsymbol{\Psi}_{t+h})$, is quite flexible, given that the weights it attaches to each of the J densities are dynamic and because it allows for time-varying error variances. We can also see that while Eq. (2) implies a Gaussian density conditional on $\boldsymbol{\beta}_{t+h}$, $\boldsymbol{x}_{t+h|t}$, and σ^2_{t+h}, when carrying out predictive inference we marginalize out the unknowns of the model, leading to a predictive density that can be highly non-Gaussian; see Eq. (3) below.

In contrast with other approaches to combining models and density forecasts, such as Bayesian model averaging (BMA), the weights on each density are restricted neither to lie between zero and one nor to sum to unity. In the case of BPS-RW, the degree of change in the weights will depend on the magnitude of the state innovation variances for these parameters: small values imply slow, smooth adjustment of the weights over time, while large values allow for bigger, sharper changes.

Two additional aspects of this parameterization of the synthesis function are worth noting before we introduce our regression-tree approach, which provides a more flexible nonparametric representation of the synthesis function.

First, as a special case, we define a static version of BPS that assumes time-invariant

weights $\boldsymbol{\beta}_\tau = \mathbf{0}_J$ for all τ but leaves $\boldsymbol{\gamma}$ unrestricted. We label this instance of BPS, which assumes the combination weights to be constant over time, "BPS-CONST."

Second, the presence of both an intercept and an error in the synthesis function means that these versions of BPS allow for model set "incompleteness" (Geweke, 2010). That is, they allow the "true" (but unknown) model not to be in $\mathcal{D}$'s model space; see, for example, Billio et al. (2013) and Aastveit et al. (2018). A conventional model combination scheme such as BMA sets both intercept and error to zero. The fact that the intercept, c_{t+h}, and error variance, σ^2_{t+h}, are both time varying provides additional flexibility when modeling the degree of model set incompleteness. Note that these specific assumptions are equivalent to embedding a popular benchmark for forecasting (especially of inflation) – the unobserved components SV (UCSV) model of Stock and Watson (2007) – within our set of now $J + 1$ density forecasts. This is also related to an alternative treatment of model set incompleteness in BPS that adds a fictitious baseline predictive density to the set of densities being synthesized (see, for instance, the discussion in Section 2.2.3 of Tallman and West, 2023).[4] In our case, this baseline predictive density comes from a UCSV model. But importantly, as when estimating a mixture density, the parameters of the UCSV density are estimated simultaneously with the weights in the synthesis function.

To carry out predictive inference, we need to compute the predictive distribution. We do so by simulation. Let y_{T+h} denote a future realization of our target variable at time $T + h$ and let $\mathcal{H}_T$ denote the set of agent densities that are available at time T but target $T + h$. The predictive density, in our case, is obtained as follows:

$$p(y_{T+h}|\mathcal{I}_T) = \int\int p(y_{T+h}|\boldsymbol{\Psi}_{T+h}, \mathcal{H}_T)\; d\boldsymbol{\Psi}\; d\mathcal{H}, \tag{3}$$

where $\mathcal{I}_T$ indicates the information set up to time T and $\boldsymbol{\Psi}_{T+h}$ are the latent states (projected forward to time $T + h$). We can simulate from (3) by simulating from the joint posterior of the agents and states, projecting the states forward to time $T + h$, and then using the synthesis function in (2) to produce a combined forecast draw. By doing so, we integrate out the unknowns of the model, and the resulting predictive density can be highly non-Gaussian and feature heavy tails, multi-modalities, and/or skewness.

[4] Diebold et al. (2023) also add a fictitious forecaster in their ECB SPF application that, like ours below, combines forecaster-level density forecasts.

2.2 Bayesian Predictive Synthesis with Regression Trees (BPS-RT)

In this paper, our proposal is to relax the restrictions in BPS-RW by considering more flexible forms of time variation in $\boldsymbol{\beta}_{t+h}$. Specifically, we use techniques from machine learning to model the dynamic evolution of the weights, $\boldsymbol{\beta}_{t+h}$, in a nonparametric manner as a function of additional weight modifiers. This treatment can be contrasted with the alternative of treating the function, α, itself nonparametrically. We follow Chipman et al. (2010) and use Bayesian additive regression trees (BART) to estimate the regression trees. BART consists of a set of priors for the tree structure and the terminal nodes (the leaf parameters) and a likelihood for data in the terminal nodes.

BPS-RT differs from existing implementations of BPS through both the hierarchical priors used on elements in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{t+h}$ and the incorporation of additional covariates into $\mathcal{D}$'s information set. These are stored in a K_γ vector $\boldsymbol{z}_j^\gamma$ and a K_β vector $\boldsymbol{z}_{jt+h|t}^\beta$, both containing additional "data" known to $\mathcal{D}$ through period t.

We postulate a nonlinear relationship between the weights and the weight modifiers through functions $\mu_j^\gamma(\boldsymbol{z}_j^\gamma)$ and $\mu_j^\beta(\boldsymbol{z}_{jt+h|t}^\beta)$ that determine the state transition equation that can be interpreted as a prior. In particular, we assume

$$\gamma_j \sim \mathcal{N}(\mu_j^\gamma(\boldsymbol{z}_j^\gamma), \tau_j^\gamma) \quad \text{and} \quad \beta_{jt+h} \sim \mathcal{N}(\mu_j^\beta(\boldsymbol{z}_{jt+h|t}^\beta), \tau_j^\beta), \tag{4}$$

where τ_j^γ and τ_j^β denote prior scaling parameters. For convenience, we define $\mu_j^\gamma := \mu_j^\gamma(\boldsymbol{z}_j^\gamma)$ and $\mu_{jt+h}^\beta := \mu_j^\beta(\boldsymbol{z}_{jt+h|t}^\beta)$. The best way to illustrate the effect the scaling parameters have on the actual estimates of the weights is to consider a reparameterization of the synthesis function. Integrating out γ_j and β_{jt+h} by plugging Eq. (4) into Eq. (2) yields

$$y_{t+h} = c_{t+h} + \sum_{j=1}^{J} \left[\underbrace{\left(\mu_j^\gamma(\boldsymbol{z}_j^\gamma) + \sqrt{\tau_j^\gamma}\nu_j^\gamma\right)}_{\gamma_j} x_{jt+h|t} + \underbrace{\left(\mu_j^\beta(\boldsymbol{z}_{jt+h|t}^\beta) + \sqrt{\tau_j^\beta}\nu_{jt+h}^\beta\right)}_{\beta_{jt+h}} x_{jt+h|t} \right] + \sigma_{t+h}u_{t+h}, \tag{5}$$

with $\nu_j^\gamma, \nu_{jt+h}^\beta \sim \mathcal{N}(0,1)$ denoting process innovations. The innovations, ν_j^γ and ν_{jt+h}^β, and the corresponding scaling terms control the degree of dispersion of the actual weights from those expected under the prior mean. If the scalings are close to zero, the posterior of γ_j and β_{jt+h} is pulled toward the prior mean and the resulting estimates will be close to $\mu_j^\gamma(\boldsymbol{z}_j^\gamma)$ and

$\mu_j^\beta(\boldsymbol{z}_{jt+h|t}^\beta)$ and so strongly depend on $\boldsymbol{z}_j^\gamma$ and $\boldsymbol{z}_{jt+h|t}^\beta$. If this is not the case, the resulting scaling parameters would be larger, so that substantial deviations from the prior means are more likely. Another feature of this representation is worth emphasizing. As opposed to a model that directly approximates the synthesis function nonparametrically, the specification in (5) introduces interaction terms of the form $\mu_j^\gamma(\boldsymbol{z}_j^\gamma) \times x_{jt+h|t}$. This specific form might reduce the risk of overfitting by introducing more structure on the space of functions that we approximate.

We approximate the prior mean functions through a sum-of-trees model with S trees:

$$\mu_j^\gamma(\boldsymbol{z}_j^\gamma) \approx \sum_{s=1}^{S} g(\boldsymbol{z}_j^\gamma | \mathcal{T}_s^\gamma, \boldsymbol{\phi}_s^\gamma) \quad \text{and} \quad \mu_{jt+h}^\beta = \mu_j^\beta(\boldsymbol{z}_{jt+h|t}^\beta) \approx \sum_{s=1}^{S} g(\boldsymbol{z}_{jt+h|t}^\beta | \mathcal{T}_s^\beta, \boldsymbol{\phi}_s^\beta), \tag{6}$$

where g denotes a tree function that is parameterized by so-called tree structures, $\mathcal{T}_s^n$, and terminal node parameters, $\boldsymbol{\phi}_s^n$, for $n \in \{\beta, \gamma\}$. The basic idea behind a single tree is that the tree structures describe sequences of disjoint sets. These sets partition the input space (determined by exogenous covariates, $\boldsymbol{z}_j^\gamma$ and $\boldsymbol{z}_{jt+h|t}^\beta$, respectively). Each of these sets is associated with a particular terminal node parameter. In our case, the terminal node parameters serve as prior expectations for the γ_js and for the β_{jt+h}s. The input space is associated with vectors of variables, $\boldsymbol{z}_j^\gamma$ and $\boldsymbol{z}_{jt+h|t}^\beta$, which we refer to as weight modifiers. Note that $\boldsymbol{z}_{jt+h|t}^\beta$ could include quantities (such as moments from the agent-specific predictive densities) that explicitly target $t+h$ but are available at time t.

There are two main justifications for our BPS-RT modeling approach. First, there is no reason to restrict attention, as in BPS-RW, to random walk specifications for the evolution of β_{t+h}. BPS-RW implies, at a given point in time, a linear relationship between y_{t+h} and $\boldsymbol{x}_{t+h|t}$. This assumption might be warranted in tranquil periods. But, in unusual times, nonlinearities could be present, and exploiting these might lead to more accurate forecasts. Our regression-tree approach allows for flexibility in the way such nonlinearities are modeled and lets the "data speak." Second, and this holds across all existing instances of BPS not just BPS-RW, an implicit assumption made is that the information set available to $\mathcal{D}$ comprises exclusively the agent-based forecast densities.[5] But, in principle, additional unmodeled information is available to $\mathcal{D}$ and might help inform evolution of the weights. In our BPS-RT approach, the weight modifiers, $\boldsymbol{z}_j^\gamma$

[5] As mentioned in footnote 2, an exception is Oelrich et al. (2023), who, when combining density forecasts using the linear opinion pool, also let the weights depend on exogenous variables. Our BPS-RT model generalizes to consider BPS combinations beyond the linear special case and to allow for nonlinearities in how the weight modifiers affect the weights.

and $\boldsymbol{z}^{\beta}_{jt+h|t}$, comprise this extra information.

These weight modifiers might include characteristics of the agents' forecasts not directly reflected in their predictive distributions or other common (to agents) factors, such as general information about the macroeconomic environment. For example, $\boldsymbol{z}^{\gamma}_{j}$ might contain summary metrics of overall past forecast performance (such as the average historical forecast performance) for each agent. Or, as noted above, $\boldsymbol{z}^{\beta}_{jt+h|t}$ might contain more granular and time-varying information, such as time-varying characteristics of the agent-specific predictive densities (say their higher moments and/or time-varying measures of past forecasting performance). We provide specific context and motivate our choice of weight modifiers in the empirical applications in Section 3.1 below.

To return to the regression tree, note that it is defined by disjoint sets that are determined by splitting rules of the form $z^{\beta}_{k,jt+h|t} \leq d_k$ or $z^{\beta}_{k,jt+h|t} > d_k$, where $z^{\beta}_{k,jt+h|t}$ is the k^{th} weight modifier for the j^{th} agent/model and d_k is a threshold parameter associated with the k^{th} effect modifier, which is estimated from the data. It is important to note, however, that any splitting rule associated with the k^{th} effect modifier is common across agents and periods (that is, it is specific neither to agent j nor to period t). Hence, the thresholds d_k and thus the tree structures do not have t or j subscripts and are common across agents/models and time. Since these splitting rules effectively govern the prior mean, μ^{β}_{jt+h}, this structure in a sense captures the notion of a pooling prior and reflects the situation that $\mathcal{D}$ decides on the weights associated to each of the different agents based on using additional factors $\boldsymbol{z}^{\gamma}_{j}$ and $\boldsymbol{z}^{\beta}_{jt+h|t}$ according to a set of common decision/splitting rules. The same structure also holds for the γ_js, with the difference that the splitting rules controlling μ^{γ}_{j} pool exclusively over the cross-section and not over time (since the γ_js are time invariant).

To see this pooling feature more clearly, consider a BPS-RT model that assumes $\boldsymbol{\beta}_{t+h} = \mathbf{0}_J$ and features only a time-variant part $\boldsymbol{\gamma}$, for which the prior mean $\boldsymbol{\mu}^{\gamma} = (\mu^{\gamma}_1, \ldots, \mu^{\gamma}_J)'$ is defined by a single tree ($S = 1$) and by a single effect modifier in z^{γ}_{j} (that is, z^{γ}_{j} is a scalar with $K_{\gamma} = 1$). In this case, the prior on γ_j can be written as

$$\gamma_j = g(z^{\gamma}_{j} | \mathcal{T}^{\gamma}_{s}, \boldsymbol{\phi}^{\gamma}_{s}) + \sqrt{\tau^{\gamma}_{j}} v_j, \quad v_j \sim \mathcal{N}(0,1). \tag{7}$$

If we now compute the difference between γ_j and γ_m for distinct agents, $j \neq m$, and assume that z^{γ}_{j} and z^{γ}_{m} are similar, in the sense that both imply the same decomposition of the

input space and are thus located in the same terminal node of the tree, we end up with

$$(\gamma_j - \gamma_m) \sim \mathcal{N}(0, \tau_j^{\gamma} + \tau_m^{\gamma}). \tag{8}$$

This equation implies that if the tree suggests that the characteristics between agents are so similar that they are grouped together in the same terminal node, the same prior mean applies and the difference between prior means will be zero. The presence of the prior scaling parameters τ_j^{γ} and τ_m^{γ} will then allow for data-based testing of whether that restriction should be strongly enforced or not. Since both prior means would coincide, setting both τ_j^{γ} and τ_m^{γ} to values close to zero would induce a clustering of γ_j and γ_m around $g(z_j^{\gamma}|\mathcal{T}_s^{\gamma}, \boldsymbol{\phi}_s^{\gamma}) = g(z_m^{\gamma}|\mathcal{T}_s^{\gamma}, \boldsymbol{\phi}_s^{\gamma})$. Hence, the choice of the prior specified on the scaling parameter τ_j^{γ} is crucial in determining the clustering behavior of BPS-RT.

Another feature of our prior is that $\mathcal{D}$ adjusts her weights on the agents' densities depending on the (common) macroeconomic environment as captured by the weight modifiers, which might include, as discussed, indicators of the state of the business cycle, measures of economic uncertainty, or deterministic trends. For example, in turbulent times, larger weights on component densities that are far from Gaussian and feature, say, heavy tails might lead to better combined density forecasts. Our approach can control for this, if supported by the data.

Note that we estimate the tree structures and the terminal parameters alongside all other unknown parameters and therefore also specify priors for them. We follow here the recommendations of Chipman et al. (2010) and discuss the remaining model and prior specification issues in detail in Appendix A. This technical appendix also describes the MCMC methods used to estimate BPS-RT. In summary, these MCMC methods are straightforward. They require a method for predictive simulation from each individual model (to draw from each agent's forecast density) and a method for drawing from the regression-tree model conditional on the individual-agent draws. For BPS-RT, the algorithm is taken directly from Chipman et al. (2010).

2.3 Illustrating BPS-RT

We now explain how BPS-RT works and allocates combination weights using an illustrative toy example. Assume that, unknown to $\mathcal{D}$, the "true" data for y_t are generated by the following

threshold model:

$$y_t = \begin{cases} \rho_1 y_{t-1} + c\rho_2 y_{t-2} + \sigma_0 \nu_t, & \text{for } t = 1, \ldots, 200 \\ c\rho_1 y_{t-1} + \rho_2 y_{t-2} + \sigma_0 \nu_t, & \text{for } t = 201, \ldots, 350, \end{cases} \tag{9}$$

where $\rho_1 = 0.8, \rho_2 = -0.8$ and $\sigma_0 = 1.2$, $y_0 = y_1 = 0$, $c = 1/100$, and $\nu_t \sim \mathcal{N}(0, 1)$.

Then, $J = 2$ agents predict y_t as follows (these forecasts are one-step-ahead, $h = 1$):

$$x_{1t} \sim y_{1t} = \mathcal{N}(\rho_1 y_{t-1}, (1 - \rho_1^2)\sigma_0^2), \tag{10}$$

$$x_{2t} \sim y_{2t} = \mathcal{N}(\rho_2 y_{t-2}, (1 - \rho_2^2)\sigma_0^2). \tag{11}$$

Both agents use forecasting methods with a different type of misspecification. The first agent's forecast is almost correctly specified for the first part of the sample, but the second agent's is substantially misspecified. In the second part of the sample this switches. We would hope that BPS-RT, when combining these two misspecified densities, would put more weight on the first agent when $t \leq 200$, then increase the weight on agent 2 when $t > 200$.

Notice that the structure of the data-generating process (DGP) implies that BPS-RW is severely misspecified, since BPS-RW implies that the combination weights on the two agents evolve smoothly over time. Our more flexible choice of synthesis function, (2), conditional on choosing appropriate effect modifiers, as we shall show, is capable of accommodating the break at $t = 200$.

We consider three variables as weight modifiers. The first is a simple deterministic time trend, $z^{\beta}_{1,jt+1|t} = t + 1$, that is common to both agents. The remaining two effect modifiers are agent-specific and measure historical forecasting performance. To capture historical point forecasting performance, we consider each agent's squared forecast error (SFE) as recursively computed at time $t-1$: $z^{\beta}_{2,jt+1|t} = (y_t - \mathbb{E}(x_{jt|t-1}))^2$ for $j = 1, 2$. Then, to measure past density forecasting performance, we consider each agent's continuous ranked probability score (CRPS).[6]

Our synthesis function is given by Eq. (2). To facilitate illustration of BPS-RT, we make some simplifying assumptions. We set the time-invariant weights $\boldsymbol{\gamma} = \mathbf{0}$ and, for the prior on $\boldsymbol{\beta}_{t+1}$, set the scaling parameters equal to zero so that the weights and prior means coincide, and we focus on the single-tree case ($S = 1$). For expositional ease, we drop corresponding sub-

[6] If F is the c.d.f. of the forecast and y the subsequent realization, then $\text{CRPS}(F, y) = \int (F(x) - \mathbf{1}_{x \geq y})^2 dx$.

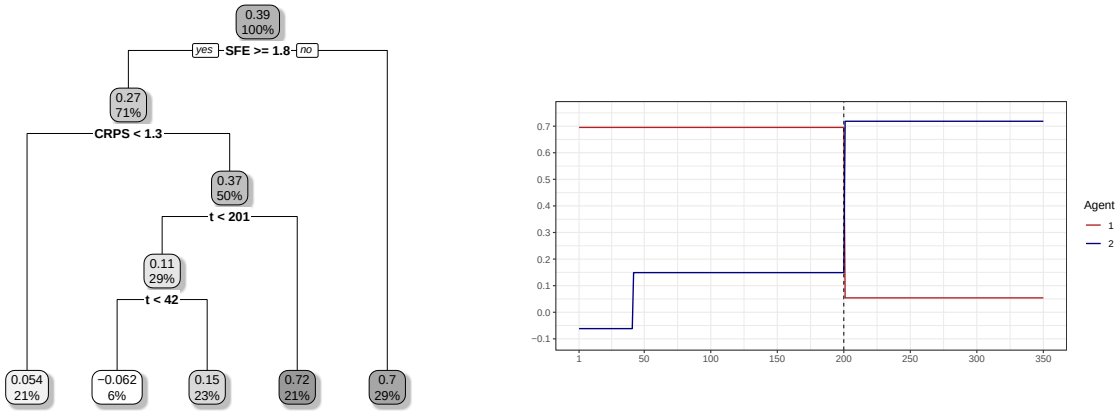
and super-scripts when there is no loss in meaning. Under these simplifying assumptions, the synthesis function, similarly to (5), reduces to

$$\alpha(y_{t+1} \mid \boldsymbol{x}_{t+1|t}, \boldsymbol{z}_{t+1|t}, \boldsymbol{\Psi}_{t+1}) = \mathcal{N}\left(y_{t+1} \mid c_{t+1} + g(\boldsymbol{z}_1|\mathcal{T}, \phi)x_{1t+1|t} + g(\boldsymbol{z}_2|\mathcal{T}, \phi)x_{2t+1|t}, \sigma^2_{t+1}\right).$$

This equation shows that with the scaling parameters set equal to zero, we end up with a BART model that assumes the weights depend nonlinearily on $\boldsymbol{z}_{t+1|t}$.

Figure 1: Illustration of BPS-RT.

(a) Decision tree of $\mathcal{D}$ (b) Combination weights over time



Notes: As the weight modifiers, we use a simple linear time trend, SFE, and CRPS. Each oval box in panel (a) indicates the terminal node parameter of a particular branch and the share (in percent) of observations belonging to this branch.

Figure 1 depicts in panel (a) the estimated tree and in panel (b) the temporal evolution of the estimated weights. We emphasize that these weights are in-sample estimates, that is, conditional on data through $T = 350$.

The tree in panel (a) can be understood as follows. Let us start at the bottom of the tree. We see five terminal nodes. Hence, we observe five groups/clusters that define the prior mean both over time and across agents. Put differently, there are five "breaks" over time and across agents in the prior mean.

How we pool is defined by the splitting rules. These are understood by turning to the top of the tree. At the root (level 0), the SFE is used as a splitting variable. The threshold parameter is 1.8 and, hence, if the SFE in $t-1$ is larger than or equal to 1.8, we move down the left branch of the tree. At the first level, the lagged CRPS shows up as the next threshold variable. If the CRPS is smaller than 1.3, we end up in a terminal node and set the weight associated with an agent that has an SFE greater than or equal to 1.8 and a CRPS smaller

than 1.3 equal to $\mathbb{E}(\beta_{jt}) = 0.054$. These conditions are fulfilled 21 percent of the time. By contrast, if the CRPS is greater than or equal to 1.3, we drop down to the second level of the tree. In this segment, time shows up as a splitting variable, and if $t \geq 201$, we assign a weight of 0.72. For $t < 201$, we introduce a further splitting rule that splits the sample once more by testing whether $t < 42$. If this is the case, a negative weight of -0.062 is applied, whereas if $42 \leq t < 201$ the weight is 0.15. If the past SFE is smaller than 1.8, we end up in the right branch of the tree and assign a weight equal to 0.7.

Hence, the tree suggests that, first and foremost, $\mathcal{D}$ selects agents according to the past performance of their forecasts, since both SFE and CRPS are identified in the estimated tree. Under our DGP, this implies that weights dynamically update if a given agent issued a poor prediction in the previous period without taking into account the past performance of her forecasts. To understand how these decision rules translate into the actual evolution of model weights, panel (b) shows the weights over time. These indicate that in the first part of the sample, Agent 1 receives substantial weight, while Agent 2 receives relatively little weight. This makes sense, given that the former is only mildly misspecified, whereas the latter features substantial model misspecification. As expected, given the structural break in the DGP, $\mathcal{D}$ now overweights the second agent, whereas the weight on Agent 1 is now much smaller.

This simple exercise illustrates how $\mathcal{D}$ incorporates additional information (time and past forecast errors in this case) to combine models. In general, though, the prior scaling parameters in BPS-RT are greater than zero, and hence, the decision tree gives rise to prior expectations that in turn inform the posterior estimates of the weights. Hence, if there is no relationship between the weights and the weight modifiers, the resulting prior variance would be large and the weights would follow a white noise process.

3 Two Macroeconomic Forecasting Applications

We investigate the performance of BPS-RT in two forecasting exercises. In the first application, we combine predictive densities of GDP growth for the euro area (EA) produced by individual professional forecasters participating in the ECB Survey of Professional Forecasters (SPF). Beyond its intrinsic interest, this data set is a good testing ground for BPS-RT because it has been used before when comparing alternative density forecast combination methods; see Diebold et al. (2023), Conflitti et al. (2015), and Chernis (2023). Second, we forecast US inflation using

a set of autoregressive distributed lag (ADL) regression models. This data set and model set has been used by Stock and Watson (2003) and Rossi and Sekhposyan (2014), the latter using a similar ADL strategy to create each of the agent's forecast densities.

These two applications differ not only geographically and in terms of target variables, but also in the number of agents and the nature of the forecast densities the agents provide. The EA GDP growth application features a relatively small number of subjective, most likely judgment-informed, forecasts (ECB, 2019) that are provided in the form of histograms (with $J = 14$). In contrast, the US inflation application uses a large number of model-based predictive densities, which are continuous and produced with distinct ADL regressions (with $J = 56$). Further details on the design of both applications are provided in the subsequent sub-sections 3.2 and 3.3. Both applications' evaluation samples cover the global financial crisis, the euro area crisis, and the COVID-19 pandemic. Taken together, these two applications enable a comprehensive assessment of BPS-RT.

3.1 BPS-RT Specifications

We experiment with several different specifications of BPS-RT to draw out how density forecast accuracy varies with the characteristics of the specific synthesis function used. In broad strokes, we look at the importance of time variation, in both weights and volatility, the number of trees, and the choice of weight modifiers. Accommodating temporal instabilities (for example, see Rossi, 2021) is important in macro-modeling, and so is a natural subject of inquiry, while the number of trees is an important aspect of specifying BART models. Being able to specify weight modifiers is an attractive feature of BPS-RT and allows the combination weights to change based on information exogenous to the individual agents but known to the BPS decision maker. Hence, this is also a key area of inquiry.

We accordingly investigate the following four specifications of BPS-RT distinguished by their choice of weight modifier(s) and whether that choice introduces cross-sectional (which we label C) or cross-sectional and time variation (which we label TC) in the combination weights seen in (2).

- **BPS-RT(C): AVG.-SCORES**: This specification uses as weight modifiers for the cross-sectional varying coefficients ($\boldsymbol{z}_j^{\gamma}$) measures of each agent's historical (ex post) forecast accuracy. Specifically, to capture past point and density forecast accuracy, it considers

model-specific averages of the mean squared forecast errors (MSEs) and the CRPSs, respectively. These averages are computed recursively to reflect information known only to $\mathcal{D}$ in real time and could help the synthesis function distinguish between "good" and "bad" forecasters. This specification assumes constant weights over a given estimation window, although the weights are updated recursively through the evaluation period.

- **BPS-RT(TC): EXO.-IND**: This specification selects as weight modifiers application-specific exogenous indicators. These vary over time but not over the cross-section, implying that they are the same for each agent. These indicators are intended to provide a signal on the state of the economy, prompting BPS-RT to reweight the individual agents while simultaneously fostering a certain degree of synchronization among them. For example, during periods of high economic uncertainty, financial stress, or elevated inflation expectations, BPS-RT may weight a subset of models more heavily. In the EA SPF application, we use the European economic policy uncertainty (EPU) index of Baker et al. (2016), available via https://www.policyuncertainty.com. In recessions, their uncertainty measure rises; so, allowing the combination weights to depend on uncertainty enables them to move with the business cycle. In the US inflation application, we consider measures of inflation expectations and financial conditions. Both variables have been considered in the inflation-at-risk literature (López-Salido and Loria, 2020). Specifically, we consider households' one-year-ahead inflation expectations from the University of Michigan survey and, as a broad measure of financial conditions, the Chicago Fed's national financial conditions index (NFCI). Both measures are available from the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org). In our empirical application, where we use a direct forecast design, we lag these exogenous indicators by the forecast horizon h to acknowledge the reality that we do not observe values for them in a future period, $t + h$, but only have information up to t. To catch any other time effects, in both applications we also consider a time trend ($t = 1, \ldots, T$).

- **BPS-RT(TC): FEATURES**: In addition to the scores discussed above, this specification considers statistical "features" of each agent's predictive density, known to $\mathcal{D}$ in real time. Specifically, we consider the first four moments of each agent's predictive density and the cross-sectional dispersion of the agents. The latter is measured by the standard deviation (at time t) across the J agents' (models') mean forecasts. Consideration of these

features allows the density combination weights, in effect, to cluster to reflect the marginal properties of the individual forecasts and their disagreement. For example, it may well be that high (ex ante) uncertainty forecasters should be weighted similarly. Besides these features, we consider historical point and density forecast performance, measured by lagged MSEs and lagged CRPSs. For this RT(TC) specification, it is noteworthy that both score measures are used in such a way that the averages of these lagged scores impact the time-invariant part of the weights $\boldsymbol{\gamma}$, while the plain lagged scores impact the weights through the time-varying part $\boldsymbol{\beta}_t$.

- **BPS-RT(TC): ALL**: This specification includes all the previously discussed weight modifiers. By looking at the weight modifiers individually, and adding features sequentially, we can assess the marginal benefit of each weight modifier.

For each of these four versions of the model, we consider models with SV and homoskedastic errors and we allow the BART specification to either have a single tree ($S = 1$), leading to a Bayesian regression-tree specification (see Chipman et al., 1998), or a large number of trees ($S = 250$), leading to BART. In traditional Bayesian implementations using trees for nonlinear regression, such as Chipman et al. (2010), it is generally found that increasing the number of trees, starting at $S = 1$, leads to an improvement in forecast performance. But this improvement tends to peter out when the number of trees gets moderately large. The conventional wisdom is that the precise choice of the number of trees is not that important, provided that too small a value is not chosen. This may not be the case in BPS, since the data may prefer to have weights that are reasonably constant over time and change only occasionally. Hence, we choose to focus on single-tree specifications and BART to model the weights in BPS. As we shall see, we find that single-tree methods tend to forecast more accurately. As benchmarks in the forecasting exercises below, we consider both BPS-CONST and BPS-RW (as defined in Section 2.1.2).

3.2 Forecasting EA Output Growth Using the Survey of Professional Forecasters

The ECB has been producing the SPF since 1999. The ECB SPF is the longest-running EA survey of macroeconomic forecasts. Each quarter, the survey elicits from a panel of professional forecasters point and probability forecasts of EA inflation and GDP growth at various horizons.[7]

[7]For a full description of the EA SPF, see Garcia (2003).

We consider the two-quarter-ahead forecasts of year-on-year EA GDP growth. On average, there are 50 responses a quarter from a survey panel of over 100 professional forecasters.

There are a couple of features of the forecaster-level density forecasts from the ECB SPF that we have to address in order to combine them. First, survey respondents provide their probability forecasts over given (fixed) ranges. That is, they produce histogram rather than continuous density forecasts. For example, in the 1999Q1 survey, forecasters were instructed to provide their probability forecasts over 10 bins. The first bin was GDP growth less than 0 percent, with the bins then increasing in intervals of 50 basis points, until the tenth bin of higher than 4 percent growth. To accommodate the discretized nature of these probability forecasts, rather than fit a continuous density to the histogram (that may or may not have a good fit), we use the histogram forecast data "as is." We do this by, within our BPS approach, drawing samples for each forecaster directly from the histograms. Details of our algorithm, which involves a Metropolis-Hastings step, are given in Appendix A.2. Our sampling approach changes over time to capture the fact that the bin definitions have been moved over time. In particular, after shocks such as the global financial crisis and COVID-19, the ECB shifted the bins to allow forecasters to say more about the probabilities in what were, prior to the survey change, the extremes of the distribution. We also have to take a stand on the open intervals at the bottom and top of the histogram. We set the end-points for the histograms equal to the outer bin plus or minus (depending on whether we are at the top or bottom of the histogram) two standard deviations of GDP growth, as estimated using the vintage of GDP data available at the time the forecast was made.

Second, forecasters enter and exit the panel. This means that the panel is unbalanced. We follow Diebold et al. (2023) in constructing the longest consistent panel possible by dropping forecasters who are regular non-responders and then filling in the occasional missing values for the remaining forecasters. Specifically, we drop forecasters who have not responded for five or more consecutive quarters. This results in a panel of 14 forecasters. Any missing forecast data for these 14 forecasters are estimated using a Normal distribution based on the unconditional distribution of GDP growth as estimated in real time.[8]

We then take these 14 forecasters' densities and carry out a recursive out-of-sample evalu-

[8]We differ from Diebold et al. (2023) in two ways. First, they interpolate missing forecasts based on historical performance. Second, we have a different number of forecasts because we use a different sample and we forecast GDP growth instead of inflation.

ation of the alternative BPS specifications over the sample 2005Q2 through 2021Q1. To do this, we first estimate the BPS combinations on a set of training samples that comprise a sequence of expanding windows of GDP and density forecast data. The GDP data used in the training sample are that vintage of GDP data available to the forecasters when they made their forecasts. The first training sample uses forecasts from the five-year period targeting GDP outturns from 1999Q3 through 2004Q2. These forecasts are taken from the surveys administered between 1999Q1 and 2003Q4. Given its publication lags and our desire to approximate the information set available at the time the SPF forecasts are publicly available, the GDP outturns required to estimate the BPS synthesis function over this training sample are taken from the 2004Q4 vintage. This estimated synthesis function then uses the 2004Q4 survey to forecast (out-of-sample) 2005Q2. The training sample and vintage of GDP data are then extended by one quarter, and forecasts are produced for 2005Q3. This process is continued until forecasts are produced for 2021Q1. This set of out-of-sample BPS density forecasts is then evaluated against GDP outturns taken from the June 9, 2021, vintage.

3.3 Forecasting US inflation Using a Set of Indicators from FRED-QD

We follow Rossi and Sekhposyan (2014) and construct density forecasts of US inflation using a set of autoregressive distributed lag (ADL) models. Each ADL model considers 1 of 27 indicators taken from the FRED-QD data set (McCracken and Ng, 2021), which is commonly used when forecasting macroeconomic aggregates such as inflation in the US. The selected indicators capture movements in assets prices, measures of real economic activity, wages and prices, and money. This rich and diverse set of economic indicators allows the ADL density forecasts of US inflation to display significant heterogeneity. Table A.1 in the Appendix provides an overview of the variables used as exogenous predictors and the transformations applied to ensure their stationarity.

We then use each of these ADL models to produce direct forecasts for quarter-on-quarter consumer price (`CPIAUCSL`) inflation one quarter ahead ($h = 1$) and one year ahead ($h = 4$). Specifically, for each indicator, x_{jt}, for $j = 1, ..., 27$, we estimate the set of ADL models:

$$\pi_{t+h} = \rho_\pi \pi_t + \alpha_\pi x_{jt} + \varepsilon_{\pi,t+h}, \quad \varepsilon_{\pi,t+h} \sim \mathcal{N}\left(0, \sigma^2_{\pi,t+h}\right), \tag{12}$$

where π_t is inflation, ρ_π is the autoregressive coefficient, and α_π denotes the coefficient related

to the j^{th} exogenous indicator.[9] We supplement these $j = 1, \ldots, 27$ models with a 28^{th} model (the AR(1) model) that sets $x_t = 0$ in Eq. (12). We also allow $\sigma^2_{\pi,t+h}$, the error variance, to be both time-varying and constant. Hence, we estimate 28 models both with and without SV, delivering, in total, a set of 56 individual models whose density forecasts we then combine using BPS. All 56 models are estimated using standard Bayesian techniques. Details are provided in Appendix A.3.

We first estimate these models on a training sample from 1970Q1 to 1989Q4. We then iterate forward using a rolling estimation window of 80 quarters to account for possible structural changes in the US economy. The first ten years of forecasts (1990Q1 to 1999Q4) are used as a training window to estimate the BPS synthesis functions. The combined forecasts are then assessed on the evaluation sample 2000Q1 to 2022Q4. This evaluation period includes distinct economic periods characterized by different inflation dynamics, including the dotcom crash, the global financial crisis, the COVID-19 period, and the post-pandemic inflationary period.

3.4 Empirical Results

We break the empirical results into three parts presented in the following three sub-sections. First, we evaluate the relative and absolute density forecast accuracy of BPS-RT. Second, we examine why BPS-RT forecasts more accurately than the benchmarks by comparing features of their forecast densities. Third, we demonstrate aspects of interpretability of BPS-RT by examining how BPS-RT can be used to understand the role of model incompleteness, agent clustering, and the time-varying importance of the different effect modifiers.

3.4.1 Forecast Accuracy

We evaluate forecast accuracy in several ways. We first evaluate the point (conditional mean) forecasts, extracted from the combined densities, using the root mean squared forecast error (RMSE) loss function. Second, we evaluate the full predictive densities. We emphasize evaluation of the predictive densities rather than the point forecasts. Since the loss functions of forecast users tend not to be quadratic – as the density forecast literature (see Aastveit et al., 2019) emphasizes – it is always important to produce and evaluate complete probabilistic forecasts. We measure the relative forecast accuracy of the forecast densities using two popular metrics: CRPS

[9] For notational ease, we do not use j subscripts to distinguish parameters in Eq. (12).

and a tail-weighted CRPS. Both are loss functions that score the density forecast according to the realization that subsequently materializes. CRPS evaluates the "whole" density, while tail-weighted CRPS focuses on accuracy in the tails (Gneiting and Ranjan, 2011).[10] We also test the absolute calibration of the combined density forecasts using the Rossi and Sekhposyan (2019) test on the probability integral transforms (PITs); and we assess the temporal stability of forecast performance using the fluctuation test of Giacomini and Rossi (2010). The results of both these tests are summarized below, with full results presented in Appendix B.

Figures 2 and 3 report the relative forecast performance of the different models in the EA GDP growth and US inflation applications, respectively, using the RMSE, CRPS, and CRPS-tails loss functions. Each row in these figures reports the relative (to the BPS-RW benchmark) performance of the four BPS-RT specifications as differentiated by whether they use a single tree or 250 trees and whether they have SV or homoskedastic errors. The four columns in the figures refer to which set of weight modifiers is used.

Looking first at the RMSE panel in Figure 2 for EA GDP growth, we see little difference between the alternative BPS-RT specifications in terms of their point forecast accuracy. The accuracy of the BPS-RT specifications also tends to be similar to that of BPS-CONST and BPS-RW, with gains/losses in general only around 3 percent. This supports the stylized fact from the forecasting literature that equal-weighted combinations of point forecasts are hard to beat (see Timmermann, 2006). Turning to US inflation (Figure 3), we do see in the RMSE panel that some of the tree-based methods now improve upon the point forecast accuracy of both benchmarks and in a manner that is statistically significant. Of particular note is the superior performance of the single-tree models, which almost always outperform the more complicated 250-tree models. We discuss this finding further below.

The CRPS panels in both Figures 2 and 3 reveal yet more of a payoff to using BPS-RT, certainly relative to BPS-RW, when we evaluate the whole density. Many of the forecast accuracy gains for BPS-RT are statistically significant. An implication of this finding is that BPS-RW's assumption that the combination weights follow a random walk is not supported by the data. But BPS-CONST, especially when BPS allows for SV, remains competitive for EA GDP growth.

The CRPS and CRPS tail results echo those under RMSE loss in concluding that single-

[10] In the empirical appendix, we follow Gneiting and Ranjan (2011) and break CRPS tails into their left and right tails. See Figures B.1 and B.2 in Appendix B.

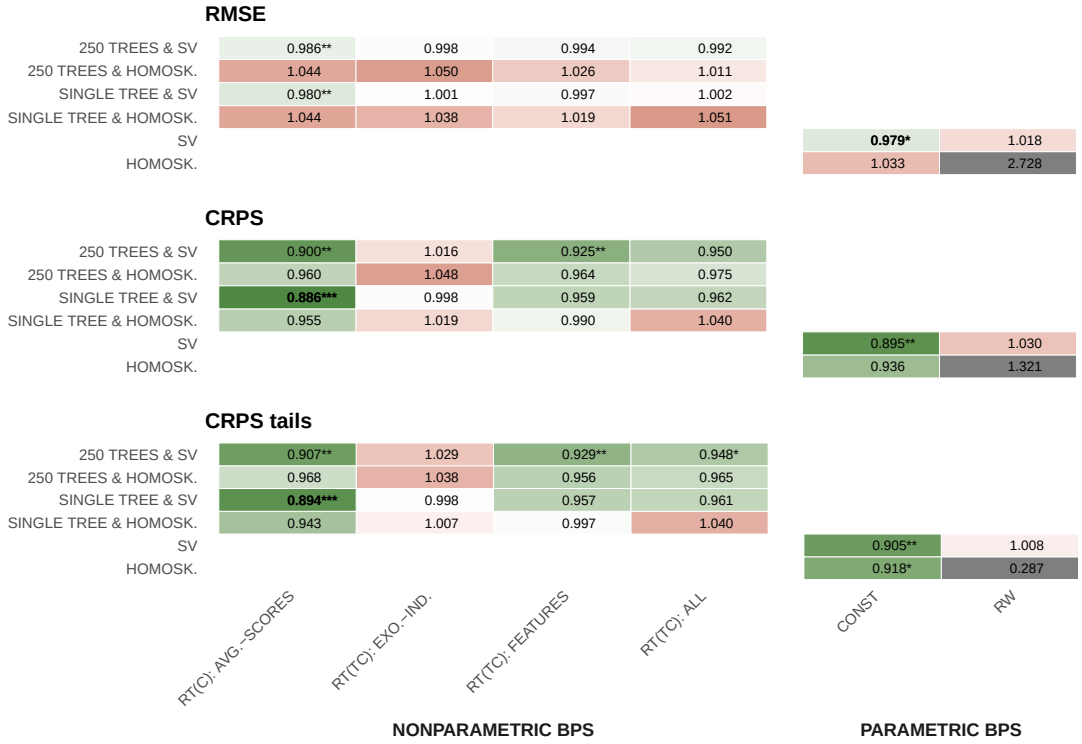
tree structures, $S = 1$, are almost always preferred to $S = 250$. The fact that a single-tree model produces more accurate forecasts contrasts with the conventional wisdom in the wider BART literature; see Chipman et al. (2010). In our case, however, we model the weights, rather than the observed outcomes, nonparametrically; hence, the implied conditional mean relation (see Equation 5) introduces more restrictions relative to a standard BART model and hence lessens the risk of overfitting.

While the benefits of allowing for SV are well established in the density forecast literature (see Clark, 2011), allowing for SV in the BPS combination does not obviously improve the density forecasts from BPS-RT. But recall, and we touch on this again below when showing that these models in fact receive higher combination weights, in the US inflation application, half of the components models themselves allow for SV.

We now focus on comparing forecast accuracy across the first four columns of both Figures 2 and 3. This comparison reveals that the choice of weight modifier does affect forecast accuracy. It is not always the case that using more weight modifiers delivers more accurate forecasts. The benefit of different modifiers varies by application and by which row (which of the four BPS-RT specifications) is consulted.

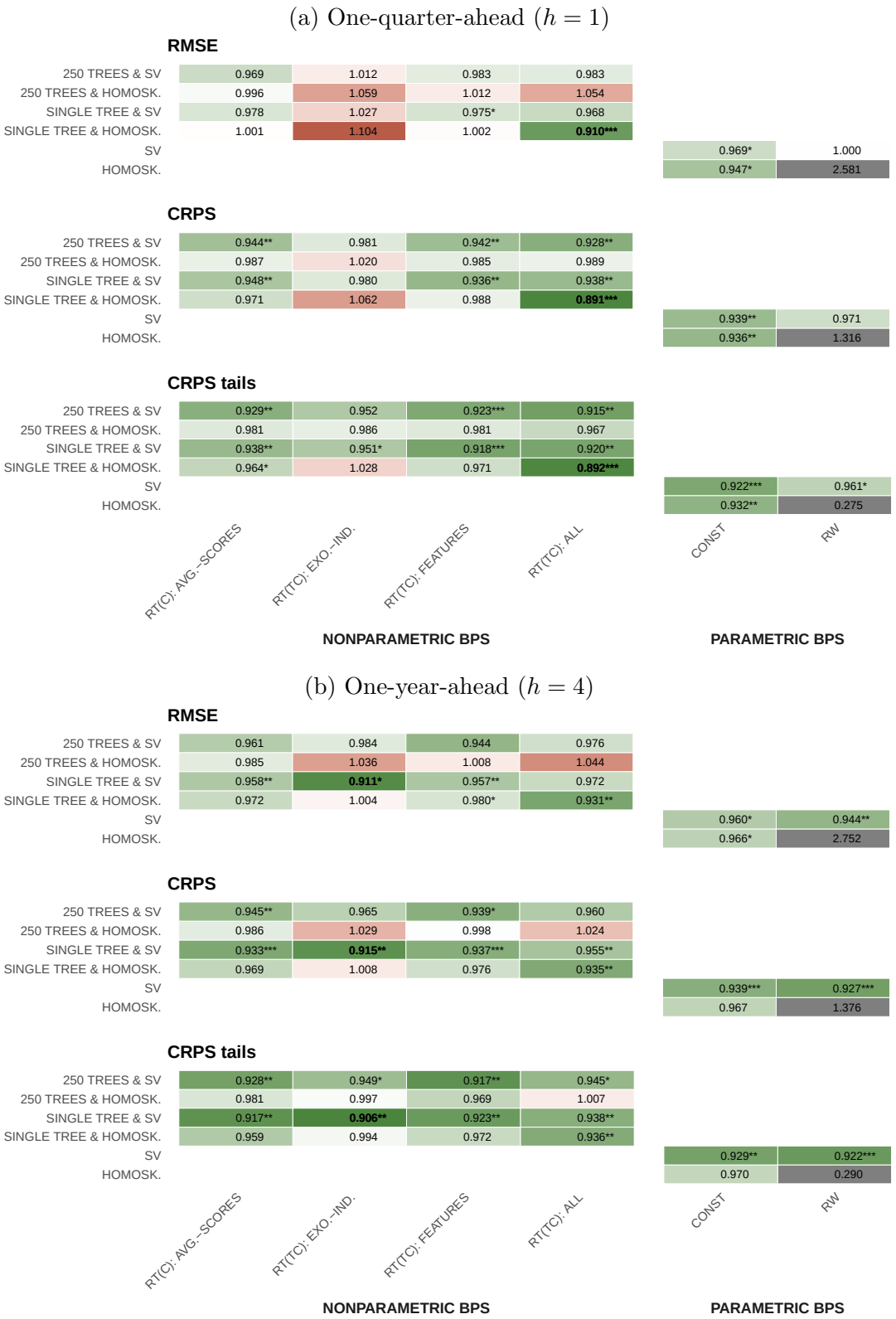
Finally, we summarize the results from both the PITs calibration tests and the fluctuation tests. These results are reported in the online appendix for space reasons. The PITs plots (see Figure B.10) show that the BPS-RT densities are well calibrated, and especially so when forecasting EA GDP growth or US inflation one-quarter-ahead. The fluctuation test of Giacomini and Rossi (2010) reveals that there is temporal variation in the relative performance (under CRPS loss) of the preferred BPS-RT model and BPW-RW. Results (see Figure B.9) indicate that the superior performance of BPS-RT in the EA application is due to better forecasting performance toward the end of the global financial crisis. For US inflation, the better accuracy of BPS-RT is explained by its more accurate density forecasts in the post-lockdown inflationary period.

Figure 2: Relative forecast accuracy: EA GDP growth.



Notes: This figure shows root mean square error (RMSE) ratios, (unweighted) continuous ranked probability score (CRPS) ratios, and a variant of quantile-weighted CRPS ratios that focuses on the tails. The gray-shaded entries give the actual scores of our benchmark (BPS-RW with homoskedastic error variances). Green-shaded entries refer to models that outperform the benchmark (with the forecast metric ratios below one), while red-shaded entries denote models that are outperformed by the benchmark (with the forecast metric ratios greater than one). The best-performing model specification by forecast metric is given in bold. Asterisks indicate statistical significance of the Diebold and Mariano (1995) test, which tests equal forecast performance for each model relative to the benchmark, at the 1 (***), 5 (**), and 10 (*) percent significance levels.

Figure 3: Relative forecast accuracy: US inflation.



Notes: This figure shows root mean square error (RMSE) ratios, (unweighted) continuous ranked probability score (CRPS) ratios, and a variant of quantile-weighted CRPS ratios that focuses on the tails. The gray-shaded entries give the actual scores of our benchmark (BPS-RW with homoskedastic error variances). Green-shaded entries refer to models that outperform the benchmark (with the forecast metric ratios below one), while red-shaded entries denote models that are outperformed by the benchmark (with the forecast metric ratios greater than one). The best-performing model specification by forecast metric is given in bold. Asterisks indicate statistical significance of the Diebold and Mariano (1995) test, which tests equal forecast performance for each model relative to the benchmark, at the 1 (***), 5 (**), and 10 (*) percent significance levels.

3.4.2 Properties of the BPS-RT Density Forecasts

In this section we examine how and why BPS-RT forecasts more accurately. We focus on the best-performing (most accurate) model in each application and compare its forecast densities to those of the benchmark model, BPS-RW.[11]

Figure 4 shows a heat map of the difference in probabilities, in intervals of 1.5 percentage point for EA GDP growth and of 1 percentage point for US inflation, between BPS-RT and BPS-RW. Green (red) shading indicates that BPS-RT adds (subtracts) probability relative to BPS-RW in that interval. This is the approach pioneered by Diebold et al. (2023) as a way of visualizing the differences between competing density forecasts.[12]

Panel (a) of Figure 4 shows that, in general, BPS-RT predictions are less dispersed than BPS-RW with more mass near the subsequent outcomes. Additionally, the BPS-RT density adds probability to low GDP growth outturns prior to the financial crisis and also forecasts higher growth than BPS-RW in both the post-global financial crisis recovery and the rebound from the COVID-19-induced recession.[13] Panels (b) and (c) of Figure 4 show the analogous plots for US inflation. Similar to panel (a), BPS-RT places more mass closer to the outturn and produces forecasts that are, in general, less disperse. Moreover, BPS-RT adjusts much more quickly to the increase in inflation post-pandemic, both one quarter ahead and one year ahead, attributing a higher probability to these outturns than BPS-RW. Consistent with the evidence in Rossi and Sekhposyan (2014) that combinations of predictive densities for US inflation appear to be approximately Gaussian, the inflation forecast densities from BPS-RT also tend to be symmetric (see Figure B.13 in the online appendix), although there is clear evidence of a spike in downside risks in 2011, a time when the Fed was engaged in quantitative easing to combat deflation threats.

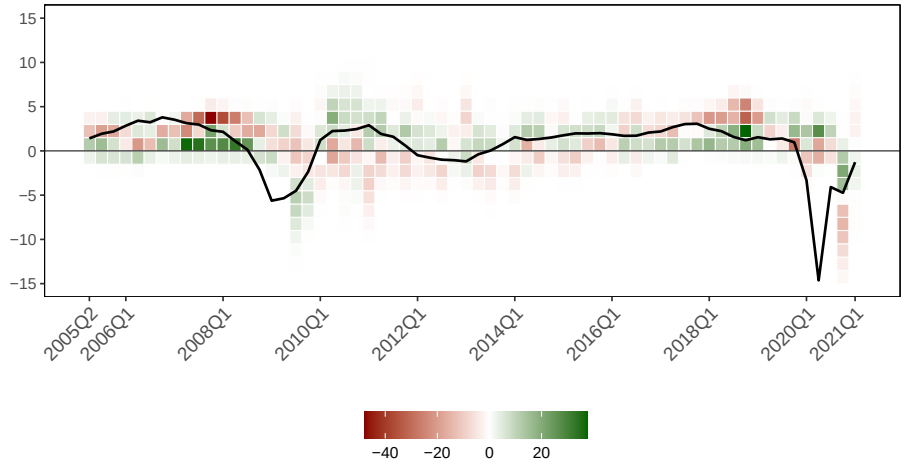
[11] As seen from Figures 2 and 3, in the EA GDP growth application, the "best" BPS-RT specification has a single tree and SV and uses average scores as effect modifiers (i.e., `RT(C): AVG.-SCORES`). For the US inflation application, the "best" BPS-RT specification has a single tree, homoskedastic errors, and the full set of weight modifiers (i.e., `RT(TC): ALL`).

[12] For an alternative but complementary visualization, Figure B.12 in Appendix B shows the temporal evolution of the underlying density forecasts from BPS-RT and the benchmark BPS-RW model over the EA and US evaluation samples.

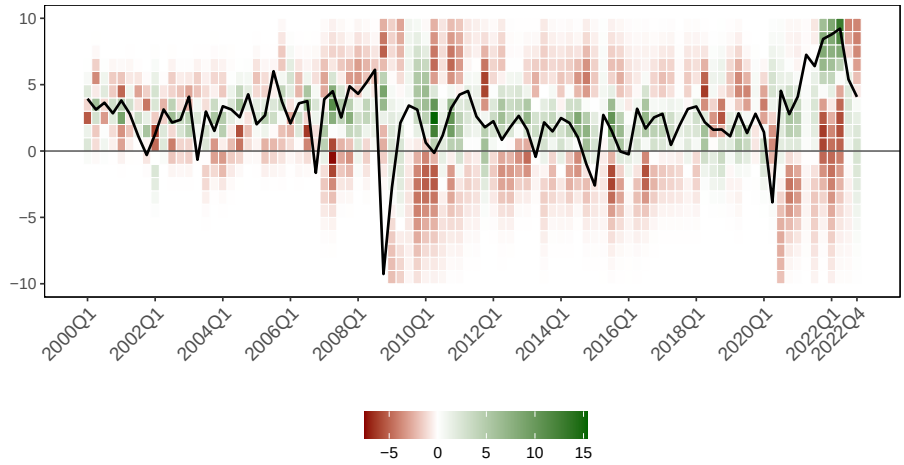
[13] As shown in Figure B.13 in the online appendix, in moving the probability mass from the centers to the left tail of the forecast density, BPS-RT captures asymmetries in the forecast densities. While there is some evidence of heightened downside risk asymmetries to GDP growth in the course of the financial crisis, consistent with the growth-at-risk literature (Adrian et al., 2019), the evidence for negative skew is stronger still during the COVID-19 pandemic.

Figure 4: Difference in probabilities between BPS-RT and BPS-RW

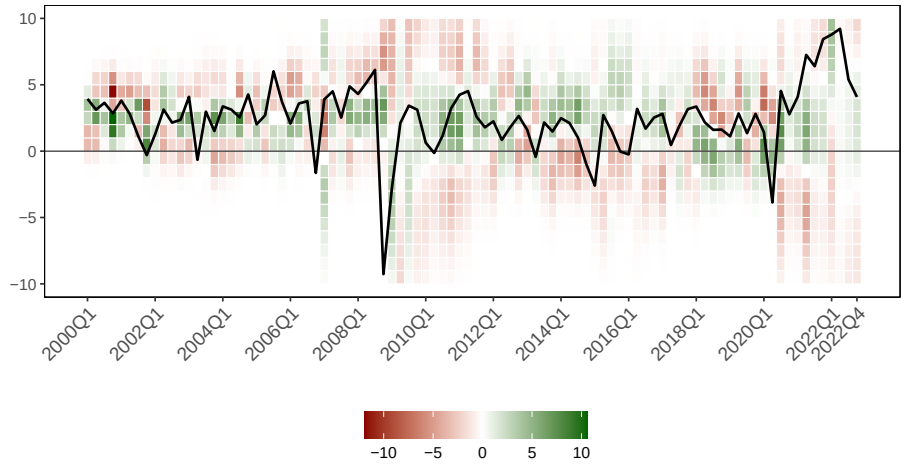
(a) EA GDP growth



(b) One-quarter-ahead US inflation ($h = 1$)



(c) One-year-ahead US inflation ($h = 4$)



Notes: This figure shows the difference in probabilities between the best-performing BPS-RT model (in terms of CRPS) and BPS-RW. We define a grid of possible values for EA GDP growth ranging from −15 percent to 15 percent with increments of 1.5 percent, while we define a grid of possible values for US inflation ranging from −10 percent to 10 percent with increments of 1 percent. Green (red)-shaded cells indicate that the best-performing model adds (subtracts) probability relative to the benchmark in the respective region.

3.4.3 Interpretation: A Deeper Dive into the Mechanics of BPS-RT for US Inflation

This section discusses how $\mathcal{D}$ can interpret the combined forecasts from BPS-RT. In so doing, we continue to focus on the preferred BPS-RT specification in the US inflation application, not least because this is where we observe greater differences across the competing combination strategies. We first show how to quantify the degree of model set incompleteness, as a way of assessing how well the agents (the J forecasting models) that BPS-RT is combining are actually able to forecast. Second, we assess the relative importance of individual weight modifiers in driving BPS-RT.

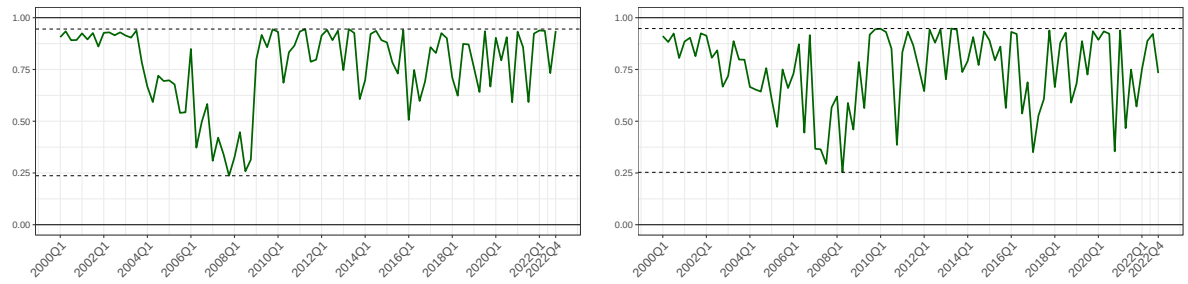
To measure model set incompleteness, we compute an R^2-type measure. This estimates the proportion of the variation in y_{t+h} that is explained by the J agents. This measure is computed, for a specific period in the evaluation sample, as the ratio between the variation in the conditional mean in Eq. (2) explained exclusively by the BPS-RT component – which is the conditional mean in Eq. (2) without the time-varying intercept c_{t+h} – and the overall variation of the target variable, y_{t+h}. R^2 values close to zero signify a high degree of model incompleteness, which means that the agents' forecasts are not informative about the target variable. Instead, the intercept and error term in the BPS synthesis function, Eq. (2), explain a large portion of the total variation. In contrast, R^2 values close to one indicate that the agents' forecasts are informative and account for the majority of the variation, implying a complete model space.

Figure 5 plots this R^2-type estimate over the evaluation sample. Given that it is computed recursively, quarter by quarter, it experiences some volatility. But Figure 5 still evidences meaningful temporal variations in the degree of model set incompleteness at both forecast horizons. We see higher model incompleteness for the one-year-ahead forecasts than for the one-quarter-ahead forecasts. This is not surprising, as producing longer-horizon forecasts is obviously more difficult. At both horizons, we see increases in model incompleteness during the period 2004–2008, a time of extreme oil price volatility as well as the global financial crisis, and in the disinflation period after the 2015 oil price shock.

Interestingly, there is no clear evidence of an increase in model incompleteness during the post-pandemic rise in inflation, reinforcing the message from Figure 4 that BPS-RT was better able to anticipate the 2021 rise in US inflation.

Figure 5: Measuring model incompleteness: US inflation

(a) One-quarter-ahead ($h = 1$) (b) One-year-ahead ($h = 4$)



Notes: This figure shows the evolution of the model incompleteness measure over time. For each quarter in the evaluation sample, this measure is computed for our preferred specification (homoskedastic `BPS-RT(TC): ALL` with a single tree) as the ratio between the variation explained exclusively by the BPS-RT part (i.e., the conditional mean without the time-varying intercept) and the total variation, which thus can be interpreted as an R^2 measure. The green solid lines represent the posterior median of this incompleteness R^2, which is bounded between zero and one. Values close to zero suggest that model incompleteness, as measured by the time-varying intercept and the error variance in Eq. (2), plays an important role, while values close to one indicate that the BPS-RT part explains most of the variation.

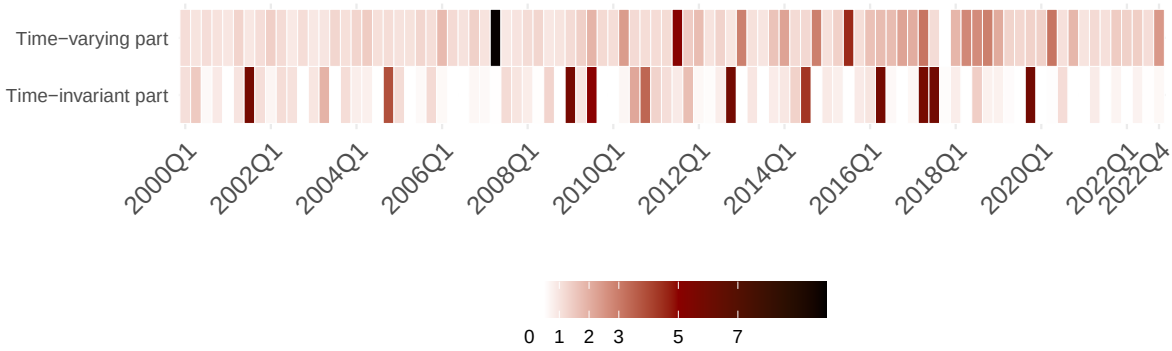
We now turn to assessing the relative importance of the individual weight modifiers in driving the density forecasts from BPS-RT. We do so by looking first at the number of tree splits and then by calculating inclusion probabilities for each weight modifier. Inclusion probabilities are calculated as the number of splits associated with the respective weight modifier divided by the total number of splits. For space reasons, we focus our discussion on Figure 6, which examines the weight modifiers for forecasting US inflation one quarter ahead. Analogous results forecasting inflation one year ahead are reported in Figure 7 and summarized below when the conclusions differ markedly from those discussed in greater detail for the one-quarter-ahead forecasts.

We start in panel (a) of Figure 6 by plotting the evolution of the total number of tree splits over the evaluation sample. This panel indicates whether variability in the combination weights comes from the time-varying (β_{jt+h}) or constant component (γ_j) of BPS-RT. Panel (a) reveals that BPS-RT tends to select a relatively small number of tree splits, especially for the time-invariant weights. Typically for γ_j, we observe that the posterior mean of the number of tree splits lies between 0.52 (lower quartile over the evaluation sample) and 1.15 (upper quartile, with a few more exceptions in the upper tail), while the average over the evaluation sample is 1.28. On the other hand, the posterior mean number of tree splits for the time-varying weights, β_{jt+h}, ranges from 1.08 to 1.68 (indicating the interquartile range) and has an average of 1.59 over the evaluation sample. To place these numbers in the context of a single-tree split on, for example, γ_j indicates that the combination weights tend to cluster around two distinct prior

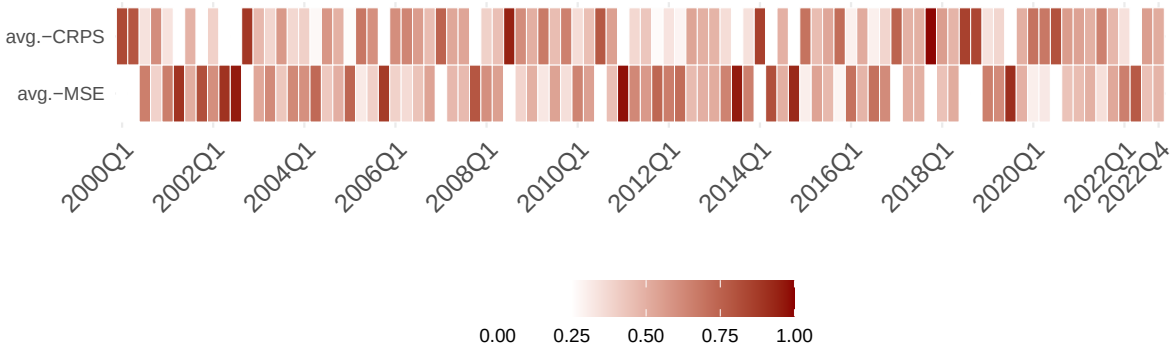
Figure 6: Number of tree splits for BPS-RT ($S = 1$) and relative importance for each weight modifier for US inflation: One-quarter-ahead forecasts

.

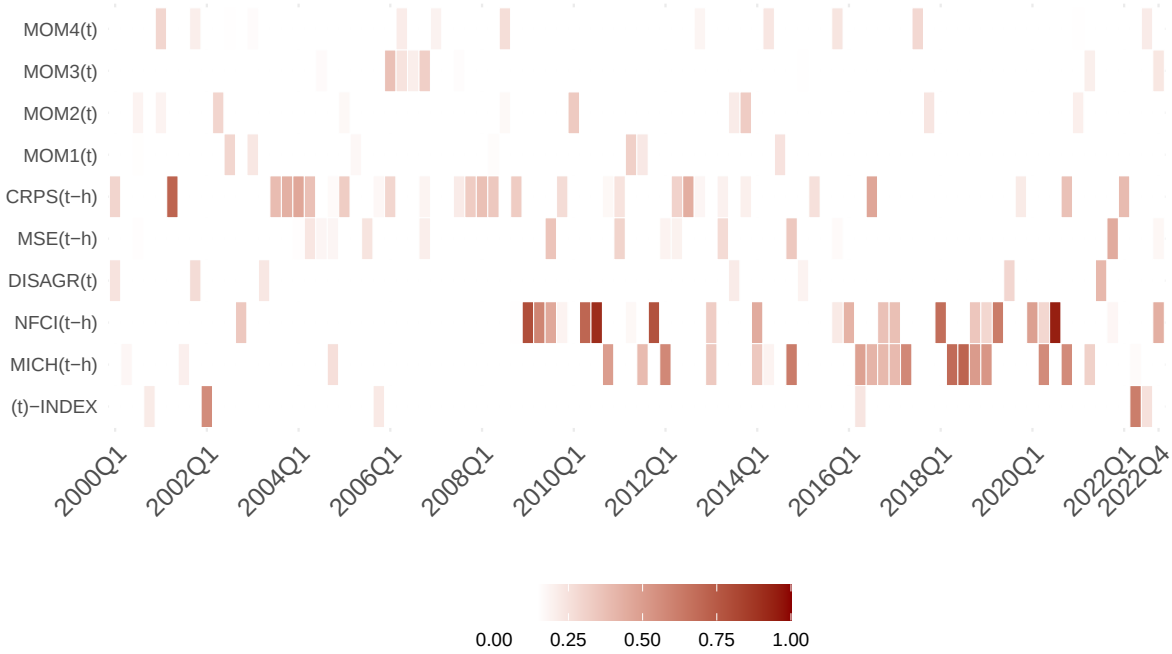
(a) **Number of (total) tree splits:** Time-invariant and time-varying weights



(b) **Time-invariant weights:** Two weight modifiers and their relative importance



(c) **Time-varying weights:** Ten weight modifiers and their relative importance



Notes: Panel (a) shows the evolution of the total number of tree splits over time, while panels (b) and (c) show the marginal importance of each weight modifier for each quarter in the evaluation sample. Relative importance is defined as the share of the total number of splits. For each quarter in the evaluation sample, we obtain the posterior mean for these measures for our preferred specification (homoskedastic `BPS-RT(TC): ALL` with a single tree). For the exogenous indicators and the MSE/CRPS scores, $(t-h)$ indicates that these measures are lagged by the forecast horizon h, while all other measures can be included contemporaneously.

means. With this in mind, we interpret the results in panel (a) as showing that the combination weights often fall into a handful of clusters that are more likely to be determined by time-specific factors. However, the number of splits is modest, so the weights are relatively stable over time. This finding is consistent with the density forecast combination literature that finds that constant weight combinations can forecast well (see, for example, Chernis, 2023).

Panels (b) and (c) of Figure 6 then show the inclusion probabilities for each of the constant and time-varying weight modifiers. Panel (b) shows the inclusion probabilities for the weight modifiers (CRPS and MSE) used to model the time-invariant combination weights. Neither CRPS nor MSE is obviously more important. Both weight modifiers receive positive and often fairly similar probabilities of inclusion. This implies that BPS-RT does partition models on the basis of their historical forecast accuracy.

Panel (c) of Figure 6 shows the importance of both time-varying weight modifiers. The first thing to notice is that there is much more sparsity in terms of the weight modifiers BPS-RT selects. In the first half of the evaluation sample, we see that features of the individual density forecasts drive the posterior inclusion probabilities. Specifically, we see that the moments of the marginal densities and CRPS, lagged by the forecast horizon h, are selected. But in the second half of the evaluation sample, we see the largest proportion of tree splits attributed to the NFCI during and immediately after recessions. The Michigan survey expectations measure also receives more weight after the financial crisis. This is evidence that nonlinear features of BPS-RT are driven by weight modifiers related to the business cycle. In other words, our BPS-RT model finds that the data support changing the combination weights abruptly with business cycle fluctuations. Finally, the time trend receives a higher weight in the post-COVID period of higher inflation seen in 2021 and 2022. This finding indicates that this inflationary episode – unprecedented within the sample – requires a substantial and rapid adjustment of the combination weights. These required weight dynamics cannot be fully captured by the business cycle weight modifiers. Instead, a time trend (or, more precisely, a time dummy) is ideal for modeling such a regime shift from low to high inflation during this exceptional period.

Finally, we summarize the properties of the posterior median estimates of the combination weights that are plotted over the evaluation sample in Appendix B. We draw out two conclusions for the combination weights estimated when forecasting US inflation one quarter ahead (see Figure B.4). First, BPS-RT places more weight on those component models with SV, espe-

cially toward the end of the evaluation sample. This corresponds to the period when BPS-RT outperforms the BPS-RW benchmark (see Figure B.8).

Second, among these SV models, only a subset receives large, in absolute value, weights. This indicates that there is some payoff, in terms of forecast accuracy, to occasionally placing a significantly higher weight on a small subset of models. Interestingly, some models get large negative weights. This amounts to short-selling those models as a "hedge" against the models with higher weights. A roughly similar pattern is seen for the one-year-ahead forecast combination weights seen in Figure B.5.[14]

While this subsection has focused on the US inflation application, we end by returning briefly to the EA GDP growth forecasting application. Figure B.3 in Appendix B, shows that the combination weights on most individual forecasters from the ECB SPF are, as anticipated given our earlier results, closer to equal than in the inflation application, where there was greater sparsity in the weights. This said, we do still see higher weights on a couple of experts (forecasters 6 and 14). We take this contrasting evidence across the two applications as empirical proof that BPS-RT is sufficiently flexible to adjust to forecasting scenarios that exhibit different dependence structures between the agents' forecasts.

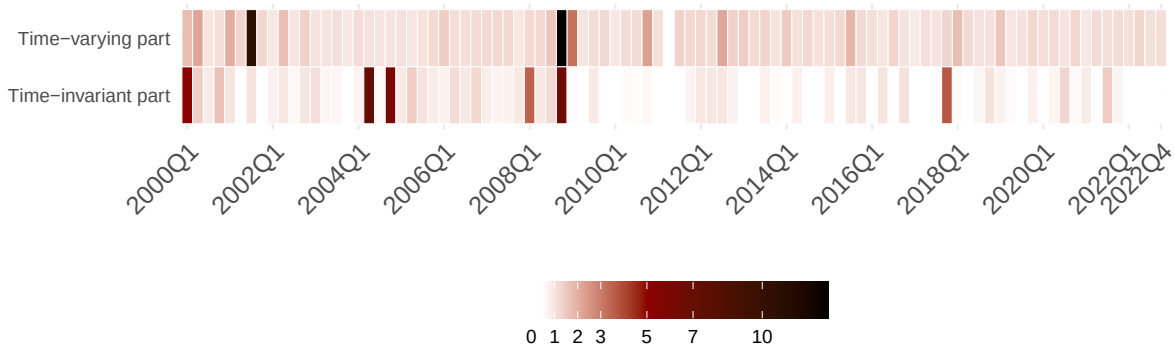
4 Conclusion

Within the general BPS framework of McAlinn and West (2019), this paper develops a method for nonparametric density forecast combination using regression trees: BPS-RT. While a handful of papers use nonparametric techniques to combine densities, ours is the first to use regression trees. In contrast to most applications of regression trees, we model the coefficients, in our case the combination weights, instead of the variables using the regression trees. We show how this aids interpretation, since the combination model remains linear in the parameters. Additionally, regression trees use covariates, or weight modifiers, to drive changes in parameters, in contrast to conventional BPS applications where model parameters follow a random walk. Taken together, our approach is flexible but retains interpretability through linearity and the use of weight modifiers. We explain how BPS-RT can be used to understand the role of model incompleteness,

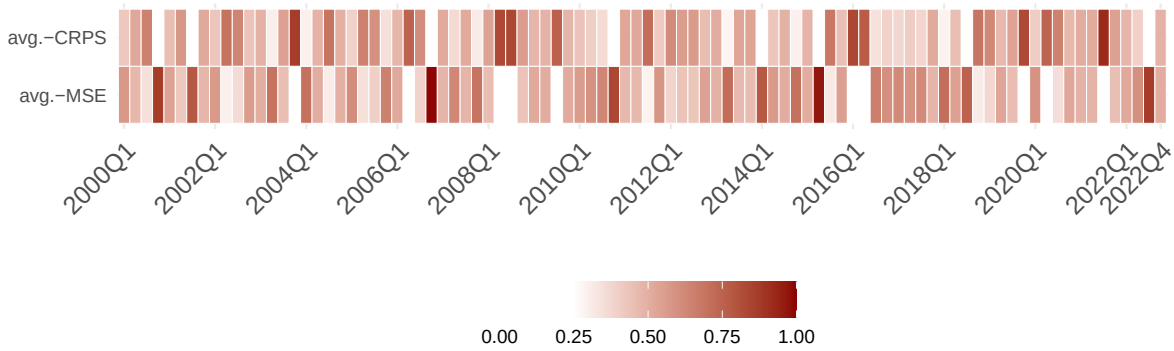
[14]Figure B.6 in the online appendix provides additional perspective on the temporal stability of the combination weights by plotting their sum over the evaluation sample. We see that when forecasting US inflation, this sum becomes negative during the global financial crisis, indicating how BPS-RT is reweighting most agents' densities in the face of temporal instabilities. The sum of the weights also spikes upward during the 2021–22 inflationary episode, again indicating how BPS-RT can quickly adapt to temporal change.

Figure 7: One-year-ahead horizon: Number of (total) tree splits for our single-tree models and relative importance for each effect modifier.

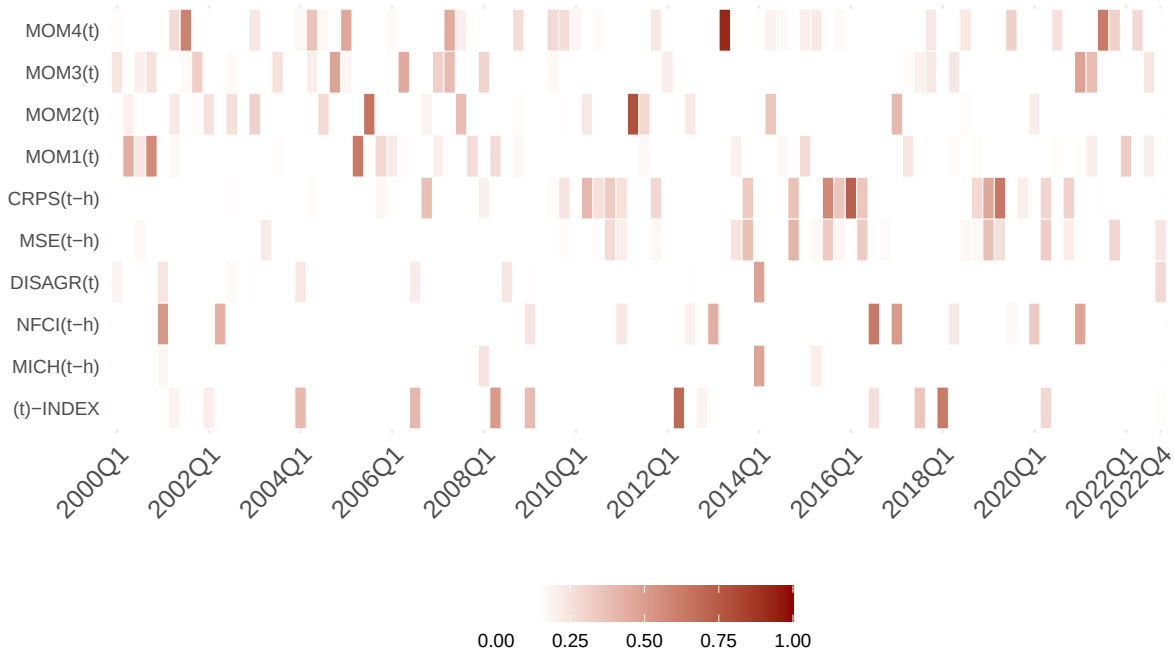
(a) **Number of (total) tree splits:** Time-invariant versus time-varying part



(b) **Time-invariant part:** Two effect modifiers and their relative importance



(c) **Time-varying part:** Ten effect modifiers and their relative importance



Notes: Panel (a) shows the evolution of the total number of tree splits over time, while panels (b) and (c) show the marginal importance of each weight modifier for each period in the evaluation sample. Relative importance is defined as the share of the total number of splits. For each period in the evaluation sample, we obtain the posterior mean for these measures for our preferred specification (homoskedastic `BPS-RT(TC): ALL` with a single tree). For the exogenous indicators and the MSE/CRPS scores, $(t-h)$ indicates that these measures are lagged by the forecast horizon h, while all other measures can be included contemporaneously.

agent (forecast) clustering, and the time-varying importance of the different weight modifiers.

We test the performance of BPS-RT in two different applications – combining model-based US inflation density forecasts and subjective histogram-based forecasts of euro area GDP growth. We find that, across both applications, BPS-RT forecasts well in terms of both relative and absolute accuracy. Interestingly, and in contrast to standard BART applications, we find that using a parsimonious single-tree specification outperforms models with more trees. Inspecting the best-performing specification, we observe that this superior performance is due to less disperse forecast densities and BPS-RT's ability to better accommodate the shocks associated with the global financial crisis (in the GDP application) and COVID-19 (in the inflation application). Our proposed measure of model set incompleteness suggests that BPS-RT is able to capture much of the post-COVID rise in inflation. Triggered by a rise in the relative importance of the time trend in determining tree splits, itself highlighting the unusual nature of this inflationary period, BPS-RT also shifts its combination weights toward component models with SV. This contrasts with the prior period of lower inflation, when the business cycle indicators were found to be more important weight modifiers.

Future lines of research could involve investigating, in other forecasting applications and contexts, the usefulness of different sets of weight modifiers and the implications for weight structure. For instance, this could draw on the ability of BPS-RT, via its choice of weight modifiers, to capture general patterns of cross-sectional dependence between competing agents' probabilistic forecasts. Additional structure could be given to the clustering by, for example, letting the combination weight on a given individual agent's density forecast depend not only on characteristics of her own forecast (such as its mean or variance) but on characteristics of the other agents' forecasts.

References

Aastveit, Knut Are, Jamie L. Cross, and Herman K. van Dijk (2023). "Quantifying time-varying forecast uncertainty and risk for the real price of oil." *Journal of Business & Economic Statistics*, 41(2), pp. 523–537. doi:10.1080/07350015.2022.2039159.

Aastveit, Knut Are, Karsten R Gerdrup, Anne Sofie Jore, and Leif Anders Thorsrud (2014). "Nowcasting GDP in real time: A density combination approach." *Journal of Business & Economic Statistics*, 32(1), pp. 48–68. doi:10.1080/07350015.2013.844155.

Aastveit, Knut Are, James Mitchell, Francesco Ravazzolo, and Herman K. van Dijk (2019). "The evolution of forecast density combinations in economics." In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press. doi:10.1093/acrefore/9780190625979.013.381.

Aastveit, Knut Are, Francesco Ravazzolo, and Herman K. van Dijk (2018). "Combined density nowcasting in an uncertain economic environment." *Journal of Business & Economic Statistics*, 36(1), pp. 131–145. doi:10.1080/07350015.2015.1137760.

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019). "Vulnerable growth." *American Economic Review*, 109(4), pp. 1263–1289. doi:10.1257/aer.20161923.

Baker, Scott, Nicholas Bloom, and Steven Davis (2016). "Measuring economic policy uncertainty." *The Quarterly Journal of Economics*, 131(4), pp. 1593–1636. doi:10.1093/qje/qjw024.

Bassetti, Federico, Roberto Casarin, and Marco Del Negro (2023). "Inference on probabilistic surveys in macroeconomics with an application to the evolution of uncertainty in the Survey of Professional Forecasters during the COVID pandemic." In Rüdiger Bachmann, Giorgio Topa, and Wilbert van der Klaauw, editors, *Handbook of Economic Expectations*, pp. 443–476. Elsevier. doi:10.1016/B978-0-12-822927-9.00023-9.

Bassetti, Federico, Roberto Casarin, and Francesco Ravazzolo (2018). "Bayesian nonparametric calibration and combination of predictive distributions." *Journal of the American Statistical Association*, 113(522), pp. 675–685. doi:10.1080/01621459.2016.1273117.

Billio, Monica, Roberto Casarin, Francesco Ravazzolo, and Herman K. van Dijk (2013). "Time-varying combinations of predictive densities using nonlinear filtering." *Journal of Econometrics*, 177(2), pp. 213–232. doi:10.1016/j.jeconom.2013.04.009.

Čapek, Jan, Jesús Crespo Cuaresma, Niko Hauzenberger, and Vlastimil Reichel (2023). "Macroeconomic forecasting in the euro area using predictive combinations of DSGE models." *International Journal of Forecasting*, 39(4), pp. 1820–1838. doi:10.1016/j.ijforecast.2022.09.002.

Carter, Chris and Robert Kohn (1994). "On Gibbs sampling for state space models." *Biometrika*, 81(3), pp. 541–553. doi:10.1093/biomet/81.3.541.

Carvalho, Carlos M, Nicholas G Polson, and James G Scott (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2), pp. 465–480. doi:10.1093/biomet/asq017.

Chernis, Tony (2023). "Combining large numbers of density predictions with Bayesian Predictive Synthesis." Staff Working Papers 23-45, Bank of Canada. doi:10.34989/swp-2023-45.

Chernis, Tony and Taylor Webley (2022). "Nowcasting Canadian GDP with density combinations." Technical Report 2022-12, Bank of Canada. doi:10.34989/sdp-2022-12.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch (1998). "Bayesian CART model search." *Journal of the American Statistical Association*, 93(443), pp. 935–948. doi:10.2307/2669832.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 4(1), pp. 266–298. doi:10.1214/09-AOAS285.

Clark, Todd E. (2011). "Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility." *Journal of Business & Economic Statistics*, 29(3), pp. 327–341. doi:10.1198/jbes.2010.09248.

Clark, Todd E., Florian Huber, Gary Koop, Massimiliano Marcellino, and Michael Pfarrhofer (2023). "Tail forecasting with multivariate Bayesian additive regression trees." *International Economic Review*, 64(3), pp. 979–1022. doi:10.1111/iere.12619.

Conflitti, Cristina, Christine De Mol, and Domenico Giannone (2015). "Optimal combination of survey forecasts." *International Journal of Forecasting*, 31(4), pp. 1096–1103. doi:10.1016/j.ijforecast.2015.03.009.

Coulombe, Philippe Goulet (2020). "The macroeconomy as a random forest." Technical report, arXiv preprint arXiv:2006.12724. doi:10.48550/arXiv.2006.12724.

Del Negro, Marco, Raiden B. Hasegawa, and Frank Schorfheide (2016). "Dynamic prediction pools: An investigation of financial frictions and forecasting performance." *Journal of Econometrics*, 192(2), pp. 391–405. doi:10.1016/j.jeconom.2016.02.006.

Deshpande, Sameer K., Ray Bai, Cecilia Balocchi, Jennifer E. Starling, and Jordan Weiss (2020). "VCBART: Bayesian trees for varying coefficients." Technical report, arXiv preprint arXiv:2003.06416. doi:10.48550/arXiv.2003.06416.

Diebold, Francis X and Roberto S Mariano (1995). "Comparing predictive accuracy." *Journal of Business & Economic Statistics*, 13(3), pp. 253–263. doi:10.1198/073500102753410444.

Diebold, Francis X., Minchul Shin, and Boyuan Zhang (2023). "On the Aggregation of Probability Assessments: Regularized Mixtures of Predictive Densities for Eurozone Inflation and Real Interest Rates." *Journal of Econometrics*, 237, p. 105321. doi:10.1016/j.jeconom.2022.06.008.

ECB (2019). "Results of the third special questionnaire for participants in the ECB Survey of Professional Forecasters." Technical report, European Central Bank. URL https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/ecb.spf201902_specialsurvey~7275f9e7e6.en.html.

Frühwirth-Schnatter, Sylvia and Helga Wagner (2010). "Stochastic model specification search for Gaussian and partial non-Gaussian state space models." *Journal of Econometrics*, 154(1), pp. 85–100. doi:10.1016/j.jeconom.2009.07.003.

Garcia, Juan A. (2003). "An introduction to the ECB's survey of professional forecasters." Occasional Paper Series 8, European Central Bank. URL https://econpapers.repec.org/paper/ecbecbops/20038.htm.

Geweke, John (2010). *Complete and Incomplete Econometric Models*. Princeton University Press, Princeton. doi:doi:10.1515/9781400835249.

Geweke, John and Gianni Amisano (2011). "Optimal prediction pools." *Journal of Econometrics*, 164(1), pp. 130–141. doi:10.1016/j.jeconom.2011.02.017.

Giacomini, Raffaella and Barbara Rossi (2010). "Forecast comparisons in unstable environments." *Journal of Applied Econometrics*, 25(4), pp. 595–620. doi:10.1002/jae.1177.

Gneiting, Tilmann and Roopesh Ranjan (2011). "Comparing density forecasts using threshold- and quantile-weighted scoring rules." *Journal of Business & Economic Statistics*, 29(3), pp. 411–422. doi:10.1198/jbes.2010.08110.

Hall, Stephen G. and James Mitchell (2007). "Combining density forecasts." *International Journal of Forecasting*, 23(1), pp. 1–13. doi:10.1016/j.ijforecast.2006.08.001.

Hauzenberger, Niko, Florian Huber, Gary Koop, and James Mitchell (2023). "Bayesian modeling of time-varying parameters using regression trees." doi:10.26509/frbc-wp-202305. Federal Reserve Bank of Cleveland WP No. 23-05.

Hauzenberger, Niko, Florian Huber, Gary Koop, and Luca Onorante (2022). "Fast and flexible Bayesian inference in time-varying parameter regression models." *Journal of Business & Economic Statistics*, 40(4), pp. 1904–1918. doi:10.1080/07350015.2021.1990772.

Huber, Florian, Gary Koop, Luca Onorante, Michael Pfarrhofer, and Josef Schreiner (2023). "Nowcasting in a pandemic using non-parametric mixed frequency VARs." *Journal of Econometrics*, 232(1), pp. 52–69. doi:10.1016/j.jeconom.2020.11.006.

Huber, Florian and Luca Rossini (2022). "Inference in Bayesian additive vector autoregressive tree models." *The Annals of Applied Statistics*, 16(1), pp. 104–123. doi:10.1214/21-AOAS1488.

Jin, Xin, John M. Maheu, and Qiao Yang (2022). "Infinite Markov pooling of predictive distributions." *Journal of Econometrics*, 228(2), pp. 302–321. doi:10.1016/j.jeconom.2021.10.010.

Kastner, Gregor and Sylvia Frühwirth-Schnatter (2014). "Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models." *Computational Statistics & Data Analysis*, 76, pp. 408–423. doi:10.1016/j.csda.2013.01.002.

Knotek, Edward S. and Saeed Zaman (2023). "Real-time density nowcasts of US inflation: A model combination approach." *International Journal of Forecasting*, 39, pp. 1736–1760. doi:10.1016/j.ijforecast.2022.04.007.

Koop, Gary and Dimitris Korobilis (2012). "Forecasting inflation using dynamic model averaging." *International Economic Review*, 53(3), pp. 867–886. doi:https://doi.org/10.1111/j.1468-2354.2012.00704.x.

Li, Li, Yanfei Kang, and Feng Li (2023). "Bayesian forecast combination using time-varying features." *International Journal of Forecasting*, 39(3), pp. 1287–1302. doi:10.1016/j.ijforecast.2022.06.002.

López-Salido, J. David and Francesca Loria (2020). "Inflation at risk." Finance and Economics Discussion Series 2020-013, Board of Governors of the Federal Reserve System (U.S.). doi:10.17016/FEDS.2020.013.

Makalic, Enes and Daniel F. Schmidt (2015). "A simple sampler for the horseshoe estimator." *IEEE Signal Processing Letters*, 23(1), pp. 179–182. doi:10.1109/LSP.2015.2503725.

McAlinn, Kenichiro, Knut Are Aastveit, Jouchi Nakajima, and Mike West (2020). "Multivariate Bayesian predictive synthesis in macroeconomic forecasting." *Journal of the American Statistical Association*, 115(531), pp. 1092–1110. doi:10.1080/01621459.2019.1660171.

McAlinn, Kenichiro and Mike West (2019). "Dynamic Bayesian predictive synthesis in time series forecasting." *Journal of Econometrics*, 210(1), pp. 155–169. doi:10.1016/j.jeconom.2018.11.010.

McCracken, Michael W. and Serena Ng (2021). "FRED-QD: A quarterly database for macroeconomic research." *Review*, 103(1), pp. 1–44. doi:10.20955/r.103.1-44.

Mitchell, James and Stephen G. Hall (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), pp. 995–1033. doi:10.1111/j.1468-0084.2005.00149.x.

Oelrich, Oscar, Mattias Villani, and Sebastian Ankargren (2023). "Local prediction pools." *Journal of Forecasting*. doi:10.1002/for.3030.

Roberts, Gareth O. and Jeffrey S. Rosenthal (2009). "Examples of adaptive MCMC." *Journal of Computational and Graphical Statistics*, 18(2), pp. 349–367. doi:10.1198/jcgs.2009.06134.

Rossi, Barbara (2021). "Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them." *Journal of Economic Literature*, 59(4), pp. 1135–90. doi:10.1257/jel.20201479.

Rossi, Barbara and Tatevik Sekhposyan (2014). "Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set." *International Journal of Forecasting*, 30(3), pp. 662–682. doi:10.1016/j.ijforecast.2013.03.005.

Rossi, Barbara and Tatevik Sekhposyan (2019). "Alternative tests for correct specification of conditional predictive densities." *Journal of Econometrics*, 208(2), pp. 638–657. doi:10.1016/j.jeconom.2018.07.008.

Stock, James H. and Mark W. Watson (2003). "Forecasting output and inflation: The role of asset prices." *Journal of Economic Literature*, 41(3), pp. 788–829. doi:10.1257/002205103322436197.

Stock, James H. and Mark W. Watson (2007). "Why has U.S. inflation become harder to forecast?" *Journal of Money, Credit and Banking*, 39(s1), pp. 3–33. doi:10.1111/j.1538-4616.2007.00014.x.

Tallman, Emily and Mike West (2023). "Bayesian predictive decision synthesis." *Journal of the Royal Statistical Society: Series B*. doi:10.1093/jrsssb/qkad109. ArXiv:2206.03815.

Timmermann, Allan (2006). "Forecast combinations." In Graham Elliott, Clive W.J. Granger, and Allan Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pp. 135–196. Elsevier. doi:10.1016/S1574-0706(05)01004-9.

Wallis, Kenneth F. (2005). "Combining density and interval forecasts: A modest proposal." *Oxford Bulletin of Economics and Statistics*, 67(s1), pp. 983–994. doi:10.1111/j.1468-0084.2005.00148.x.

West, Mike (1992). "Modelling agent forecast distributions." *Journal of the Royal Statistical Society: Series B*, 54(2), pp. 553–567. doi:10.1111/j.2517-6161.1992.tb01896.x.

West, Mike and Jo Crosse (1992). "Modelling probabilistic agent opinion." *Journal of the Royal Statistical Society: Series B*, 54(1), pp. 285–299. doi:10.1111/j.2517-6161.1992.tb01882.x.

Online Appendix

A Technical Appendix: Bayesian Inference

Before we start discussing the modeling choices, and the prior and the posterior sampler in detail, we introduce a bit of additional notation to simplify the exposition.

We can rewrite Eq. (2) as a standard TVP regression:

$$y_t = c_t + \boldsymbol{\gamma}' \boldsymbol{x}_{t|t-h} + \boldsymbol{\beta}_t' \boldsymbol{x}_{t|t-h} + \sigma_t \nu_t, \quad \nu_t \sim \mathcal{N}(0,1). \tag{A.1}$$

The priors on $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_t$ can be written in the form of multivariate Gaussian distributions:

$$\boldsymbol{\gamma} \sim \mathcal{N}(\mu^{\gamma}(\boldsymbol{Z}^{\gamma}), \underline{\boldsymbol{V}}^{\gamma}),$$
$$\boldsymbol{\beta}_t \sim \mathcal{N}(\mu^{\beta}(\boldsymbol{z}^{\beta}_{t|t-h}), \underline{\boldsymbol{V}}^{\beta}).$$

Here, $\mu^{\gamma}(\boldsymbol{Z}^{\gamma})$ and $\mu^{\beta}(\boldsymbol{Z}^{\beta}_{t|t-h})$ are both prior mean functions of dimension J and $\underline{\boldsymbol{V}}^{n} = \text{diag}(\tau_1^n, \dots, \tau_J^n)$ for $n \in \{\gamma, \beta\}$. The weight modifiers are stored, respectively, in a $(J \times K_{\gamma})$ matrix $\boldsymbol{Z}^{\gamma}$ with typical row $\boldsymbol{z}_j^{\gamma}$ and in a $(JT \times K_{\beta})$ matrix $\boldsymbol{Z}^{\beta} = (\boldsymbol{z}_1^{\beta}, \dots, \boldsymbol{z}_T^{\beta})'$, with $\boldsymbol{z}_t^{\beta} = (\boldsymbol{z}^{\beta}_{1t|t-h}, \dots, \boldsymbol{z}^{\beta}_{Jt|t-h})$, $t = 1, \dots, T$.

Notice that the prior on stacked $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_T')$ can be written as a $JT-$dimensional Gaussian distribution:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mu^{\beta}(\boldsymbol{Z}^{\beta}), \boldsymbol{I}_T \otimes \underline{\boldsymbol{V}}^{\beta}),$$

where the prior mean function $\mu^{\beta}(\boldsymbol{Z}^{\beta})$ is now also of dimension JT.

A.1 Additional Details about Our Modeling Choices and Priors

In this sub-section we provide additional details about our hierarchical prior setup used for $\boldsymbol{\gamma}$ and the time-varying part, $\boldsymbol{\beta}_t$, of the weights. In both cases, the prior mean functions, $\mu^{\gamma}(\boldsymbol{Z}^{\gamma})$ and $\mu^{\beta}(\boldsymbol{z}_t^{\beta})$, are approximated by tree functions (see Eq. 6), while the prior variances – which define the degree of shrinkage toward these prior means – are modeled with a horseshoe (HS, Carvalho et al., 2010) prior. In addition, we sketch the law of motion and modeling choices for the time-varying intercept and time-varying variances in Eq. (A.1), both of which capture the idea of model incompleteness.

Tree functions to approximate the prior mean. We closely follow here the suggestions of the Bayesian additive regression tree (BART) literature (Chipman et al., 1998, 2010) and use a similar prior setup for our tree structures $\mathcal{T}_s^n$ and terminal node parameters $\boldsymbol{\phi}_s^n$ for $n \in \{\beta, \gamma\}$. To generate the tree function, Chipman et al. (1998) and Chipman et al. (2010) suggest using a stochastic process of the following form:

1. **Prior on the tree structure $\mathcal{T}_s^n$.** Impose a decreasing probability of growing more complex trees and that a terminal node is non-terminal. This probability is assumed to be

$$\frac{c_0}{(1+\vartheta)^{c_1}},$$

for a particular terminal node at depth ϑ, with the hyperparameters $c_0 = 0.95$ and $c_1 = 2$ being two values that have been shown to be reasonable choices in much of the literature using Bayesian (additive) tree models. Chipman et al. (2010) show that this choice works well even for single-tree models. Moreover, for each splitting rule at each node, Chipman et al. (2010) propose a prior that is agnostic about the choice of the specific splitting variable and propose a natural default choice, which is to use a uniform prior on the splitting variables, treating each variable as equally likely to be used in a splitting rule.

2. **Prior on the terminal node parameters $\boldsymbol{\phi}_s^n$.** We use a Gaussian prior for the terminal node parameters. For a typical element in $\boldsymbol{\phi}_s^n$, that is

$$\phi_{j,s}^n \sim \mathcal{N}(0, c_2/S),$$

where c_2 refers to a shrinkage parameter and S to the number of trees. It is worth noting that – to avoid overfitting – the prior variances for these terminal parameters are scaled down by the number of trees and become tighter, so that each individual tree explains only a tiny fraction within the additive sum-of-tree function.

Shrinkage toward the prior mean through the horseshoe prior. The horseshoe prior amounts to setting the scaling parameters as follows:

$$\tau_j^n = \lambda^n \psi_j^n, \quad \lambda^n \sim \mathcal{C}^+(0,1), \quad \psi_j^n \sim \mathcal{C}^+(0,1), \text{ for } n \in \{\gamma, \beta\},$$

with $C^+(0,1)$ denoting the half-Cauchy distribution. The key feature of this prior is that λ^n serves as a global shrinkage parameter that pulls all weights toward the prior mean, whereas ψ_j^n allows for agent-specific deviations from this common pattern. Another representation of this prior, which simplifies posterior sampling enormously, is based on introducing inverse Gamma distributed auxiliary variables (see Makalic and Schmidt, 2015):

$$\lambda^n|\zeta^n \sim \mathcal{G}^{-1}(1/2, 1/\varphi^n), \quad \varphi^n \sim \mathcal{G}^{-1}(1/2, 1),$$
$$\psi_j^n|\varpi_j^n \sim \mathcal{G}^{-1}(1/2, 1/\varpi_j^n), \quad \varpi_j^n \sim \mathcal{G}^{-1}(1/2, 1).$$

This representation is convenient since – when combined with the likelihood – it gives rise to a simple Gibbs sampling step that involves only inverse Gamma full conditionals (see Sub-section A.2 below).

Controlling for model incompleteness. A time-varying intercept c_t and time-varying variances σ_t^2 both control for model incompleteness. The fact that both are potentially time-varying gives additional flexibility in the degree of model set incompleteness (as outlined in Sub-section 2.1.2). The time-varying intercept follows a random walk (RW) law of motion with the state equation given by

$$c_t = c_{t-1} + \eta_{c,t}, \quad \eta_{c,t} \sim \mathcal{N}(0, \sigma_c^2),$$

where σ_c^2 denotes the state innovation variance. To discipline c_t, we use a relatively tight Gamma prior on σ_c^2 and strongly push the state innovation variance toward a small positive value close to zero.

The error variances σ_t^2 in Eq. (1) can be time-varying or constant. The time-varying case is given by

$$\varsigma_t = \mu_\varsigma + \rho_\varsigma(\varsigma_{t-1} - \mu_\varsigma) + \eta_{\varsigma,t}, \quad \eta_{\varsigma,t} \sim \mathcal{N}(0, \sigma_\varsigma^2), \tag{A.2}$$

with μ_ς denoting the unconditional mean, ρ_ς the persistence parameter, and σ_ς^2 the state innovation variance of the log-volatility process. For SV we follow Kastner and Frühwirth-Schnatter (2014) and assume a Gaussian prior on $\mu_\varsigma \sim \mathcal{N}(0, 10^2)$, a (transformed) Beta prior on $(\rho_\varsigma + 1)/2 \sim \mathcal{B}(5, 1.5)$, and a Gamma prior on $\sigma_\varsigma^2 \sim \mathcal{G}(0.5, 0.5)$. Moreover, for the case of

homoskedastic errors, we assume an inverse Gamma prior on $\sigma^2 \sim i\mathcal{G}(0.01, 0.01)$.

A.2 Posterior Simulation

The prior discussed in the previous section can be combined with the likelihood to derive the full posterior over all unknown quantities in our model. Since this joint density is untractable, we use Markov chain Monte Carlo (MCMC) methods to carry out posterior simulation. In what follows, we let $\bullet$ be generic notation that indicates that we condition on all other parameters/states of the model.

We start the discussion of our posterior sampler by first describing how we estimate the latent quantities that enter the synthesis function. This includes the static and dynamic weights, the error variances, the latent trend components, and the agent-specific forecasts.

Sampling from $\{p(c_t|\bullet)\}_{t=1}^T$. We sample the full history of the random walk intercept term conditional on all other unknowns in the model using a forward-filtering backward-sampling (FFBS) step (Carter and Kohn, 1994). This is achieved by noting that

$$\underbrace{y_t - \boldsymbol{\gamma}'\boldsymbol{x}_{t|t-h} - \boldsymbol{\beta}_t'\boldsymbol{x}_{t|t-h}}_{y_t^{\gamma,\beta}} = c_t + \sigma_t\nu_t \tag{A.3}$$

is a standard unobserved components model with heteroskedastic shocks.

Sampling from $p(\boldsymbol{\gamma}|\bullet)$. The time-invariant weights are sampled from a J-dimensional Gaussian full conditional posterior distribution,

$$\boldsymbol{\gamma}|\bullet \sim \mathcal{N}(\overline{\boldsymbol{\gamma}}, \overline{\boldsymbol{V}}^{\gamma}), \tag{A.4}$$

with moments given by

$$\begin{aligned}
\overline{\boldsymbol{V}}^{\gamma} &= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + (\underline{\boldsymbol{V}}^{\gamma})^{-1})^{-1}, \\
\overline{\boldsymbol{\gamma}} &= \overline{\boldsymbol{V}}^{\gamma}\left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}^{\beta,c} + (\underline{\boldsymbol{V}}^{\gamma})^{-1}\mu^{\gamma}(\boldsymbol{Z}^{\gamma})\right),
\end{aligned}$$

where $\boldsymbol{X}$ is a $T \times J$ matrix with t^{th} row $\boldsymbol{x}_{t|t-h}'$, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2)$ and $\boldsymbol{y}^{\beta,c}$ is a T-dimensional vector with typical element $y_t - \boldsymbol{\beta}_t'\boldsymbol{x}_{t|t-h} - c_t$.

Sampling $p(\boldsymbol{\beta}|\bullet)$. The dynamic regression coefficients are simulated by writing the model

in static form. The static form of the model reads:

$$\boldsymbol{y}^{\gamma,c} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{\nu}, \quad \boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{0}_T, \boldsymbol{\Sigma}), \tag{A.5}$$

where $\boldsymbol{y}^{\gamma,c}$ is $T \times 1$ and has typical element $y_t - \boldsymbol{\gamma}'\boldsymbol{x}_{t|t-h} - c_t$ and $\boldsymbol{W}$ is a $T \times TJ$-dimensional block diagonal matrix with $\boldsymbol{W} = \text{bdiag}(\boldsymbol{x}'_{1|1-h}, \ldots, \boldsymbol{x}'_{T|T-h})$.[15] Under this static representation, the posterior of $\boldsymbol{\beta}$ takes a standard form and is multivariate Gaussian:

$$\boldsymbol{\beta}|\bullet \sim \mathcal{N}(\overline{\boldsymbol{\beta}}, \overline{\boldsymbol{V}}^{\beta}), \tag{A.6}$$

with posterior covariance matrix and mean vector given by, respectively,

$$\begin{aligned}
\overline{\boldsymbol{V}}^{\beta} &= \left(\boldsymbol{W}'\boldsymbol{\Sigma}^{-1}\boldsymbol{W} + (\boldsymbol{I}_T \otimes \underline{\boldsymbol{V}}^{\beta})^{-1}\right)^{-1}, \\
\overline{\boldsymbol{\beta}} &= \overline{\boldsymbol{V}}^{\beta}\left(\boldsymbol{W}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}^{\gamma,c} + (\boldsymbol{I}_T \otimes \underline{\boldsymbol{V}}^{\beta})^{-1}\mu^{\beta}(\boldsymbol{Z}^{\beta})\right).
\end{aligned}$$

This distribution is high dimensional even for moderate values of J, and we thus use the efficient sampler outlined in Hauzenberger et al. (2022).

Sampling from $p(\sigma_1^2, \ldots, \sigma_T^2|\bullet)$. We sample the log volatilities and associated state equation parameters using the algorithm outlined in Kastner and Frühwirth-Schnatter (2014). This step is implemented in the `R` package stochvol.

Sampling from $\{p(\boldsymbol{x}_{t|t-h}|\bullet)\}_{t=1}^T$. We draw from $\{p(\boldsymbol{x}_{t|t-h}|\bullet)\}_{t=1}^T$ on a t-by-t basis. The time t full conditional posterior of $\boldsymbol{x}_t$ is given by

$$p(\boldsymbol{x}_{t|t-h}|\bullet) \propto \mathcal{N}(y_t|c_t + \boldsymbol{\gamma}'\boldsymbol{x}_{t|t-h} + \boldsymbol{\beta}_t'\boldsymbol{x}_{t|t-h}, \sigma_t^2) \prod_{j=1:J} \pi_{jt}(x_{jt|t-h}),$$

which, unless the agent densities $\pi_{jt}(x_{jt|t-h})$ are Gaussian, takes no well-known form. In our applications, the agent densities do not have analytical representations. For example, the ECB-SPF elicits histograms from survey respondents and in the US application the available forecasts are predictive draws based on model-specific Gibbs samplers. Accordingly, we sample $\boldsymbol{x}_t$ using an adaptive Metropolis Hastings step (see, e.g., Roberts and Rosenthal, 2009). This step proposes

[15] Observations $-h, \ldots, 0$ refer to a part of the sample that we use to initialize our models.

$\boldsymbol{x}^*_{t|t-h}$ from a mixture of Gaussian distributions:

$$\boldsymbol{x}^*_{t|t-h} \sim (1-\kappa)\mathcal{N}(\boldsymbol{x}_{t|t-h}, (2.38)^2\hat{\boldsymbol{Q}}_{tm}/J) + \kappa\mathcal{N}(\boldsymbol{x}_{t|t-h}, (0.1)^2\boldsymbol{I}_J/J), \tag{A.7}$$

where $\kappa = 0.05$ is a small constant and $\hat{\boldsymbol{Q}}_{tm}/J$ is the empirical covariance matrix of the target distribution based on the first m draws. Since this algorithm learns the proposal, it can quickly adjust to cases where the agent densities are non-Gaussian, feature multiple modes, or are heavy tailed.

Next, we discuss the steps involved in sampling the parameters of the priors on the weights.

Sampling from $p(\mathcal{T}^n_1, \ldots, \mathcal{T}^n_S, \boldsymbol{\phi}^n_1, \ldots, \boldsymbol{\phi}^n_S|\bullet)$ for $n \in \{\gamma, \beta\}$. We sample the tree structures and the terminal node parameters using the algorithm proposed in Chipman et al. (2010). This algorithm is applicable since, conditional on $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, the corresponding priors can be interpreted as regression models. For instance, in the case of $\boldsymbol{\beta}$, notice that

$$\boldsymbol{\beta}_{jt} = \sum_{s=1}^{S} g(\boldsymbol{z}^{\beta}_{jt}|\mathcal{T}^{\beta}_s, \boldsymbol{\phi}^{\beta}_s) + \tau^{\beta}_j \nu_j t, \quad \nu_{jt} \sim \mathcal{N}(0,1), \tag{A.8}$$

which, in stacked form, can be written as

$$\boldsymbol{\beta} = \sum_{s=1}^{S} g(\boldsymbol{Z}^{\beta}|\mathcal{T}^{\beta}_s, \boldsymbol{\phi}^{\beta}_s) + \boldsymbol{r}, \quad \boldsymbol{r} \sim \mathcal{N}(\boldsymbol{0}_{TJ}, \boldsymbol{I}_T \otimes \underline{\boldsymbol{V}}^{\beta}). \tag{A.9}$$

Equation A.9 is a standard BART regression with latent responses and heteroskedastic errors. For $\boldsymbol{\gamma}$, a similar regression representation can be derived.

Sampling from $p(\tau^n_1, \ldots, \tau^n_J|\bullet)$ for $n \in \{\gamma, \beta\}$. The scaling parameters are obtained using the algorithm described in Makalic and Schmidt (2015). This algorithm involves only inverse Gamma distributions and, for brevity, we do not discuss them in detail here.

These steps form our MCMC algorithm. In all our empirical work, we iteratively sample from the different full conditionals to obtain draws from the joint posterior of the coefficients and the latent states. Based on these draws, we back out the predictive distribution as described in the main text through Monte Carlo integration. That is, after obtaining a draw from the posterior, we use this draw to forecast y_{t^*}. This is done for every draw, leading to a posterior distribution over future values y_{t^*}. In all our empirical work, we repeat this $12,500$ times and discard the initial $2,500$ draws as a burn-in. We subsequently keep every second draw, yielding

a total of $5,000$ draws from the joint posterior distribution.

A.3 ADL Model Estimation

Our US inflation forecasting exercise involves Bayesian estimation of ADL models involving different explanatory variables. Table A.1 provides an overview of the 27 variables each used as an exogenous predictor in Eq. (12) and also highlights (in bold typeface) the target variable.

Table A.1: List of variables used for the autoregressive distributed lag (ADL) specifications.

Mnemonic	Description	Transformation
GDPC1	Real Gross Domestic Product	$100 \times \Delta \log$
PCECC96	Real Personal Consumption Expenditures	$100 \times \Delta \log$
FPIx	Real Private Fixed Investment	$100 \times \Delta \log$
GCEC1	Real Government Consumption Expenditures and Gross Investment	$100 \times \Delta \log$
INDPRO	Total index Industrial Production Index	$100 \times \Delta \log$
CUMFNS	Capacity Utilization: Manufacturing (SIC)	none
PAYEMS	Emp:Nonfarm All Employees: Total nonfarm	$100 \times \Delta \log$
CE16OV	Civilian Employment	$100 \times \Delta \log$
UNRATE	Civilian Unemployment Rate	Δ
AWHMAN	Average Weekly Hours of Production and Nonsupervisory Employees: Manufacturing Hours	none
CES0600000007	Average Weekly Hours of Production and Nonsupervisory Employees: Goods-Producing	Δ
CLAIMSx	Initial Claims	$100 \times \Delta \log$
GDPCTPI	Gross Domestic Product: Chain-type Price Index	$100 \times \Delta \log$
CPIAUCSL	**Consumer Price Index for All Urban Consumers**	$100 \times \Delta \log$
PPIACO	Producer Price Index for All Commodities	$100 \times \Delta \log$
WPSID61	Producer Price Index by Commodity Intermediate Materials: Supplies & Components	$100 \times \Delta \log$
WPSID62	Producer Price Index: Crude Materials for Further Processing	$100 \times \Delta \log$
COMPRNFB	Nonfarm Business Sector: Real Compensation Per Hour	$100 \times \Delta \log$
ULCNFB	Nonfarm Business Sector: Unit Labor Cost	$100 \times \Delta \log$
CES0600000008	Average Hourly Earnings of Production and Nonsupervisory Employees	$100 \times \Delta \log$
FEDFUNDS	Effective Federal Funds Rate	Δ
BAA10YM	Moodys Seasoned BAA Corporate Bond Yield Relative to Yield on 10-Year Treasury	none
GS10TB3Mx	10-Year Treasury Constant Maturity Minus 3-Month Treasury Bill, secondary market	none
CPF3MTB3Mx	3-Month Commercial Paper Minus 3-Month Treasury Bill, secondary market	none
M2REAL	Real M2 Money Stock	$100 \times \Delta \log$
BUSLOANSx	Real Commercial and Industrial Loans, All Commercial Banks	$100 \times \Delta \log$
CONSUMERx	Real Consumer Loans at All Commercial Banks	$100 \times \Delta \log$
S.P.500	S&Ps Common Stock Price Index	$100 \times \Delta \log$

Notes: The variable in bold refers to the target inflation series.

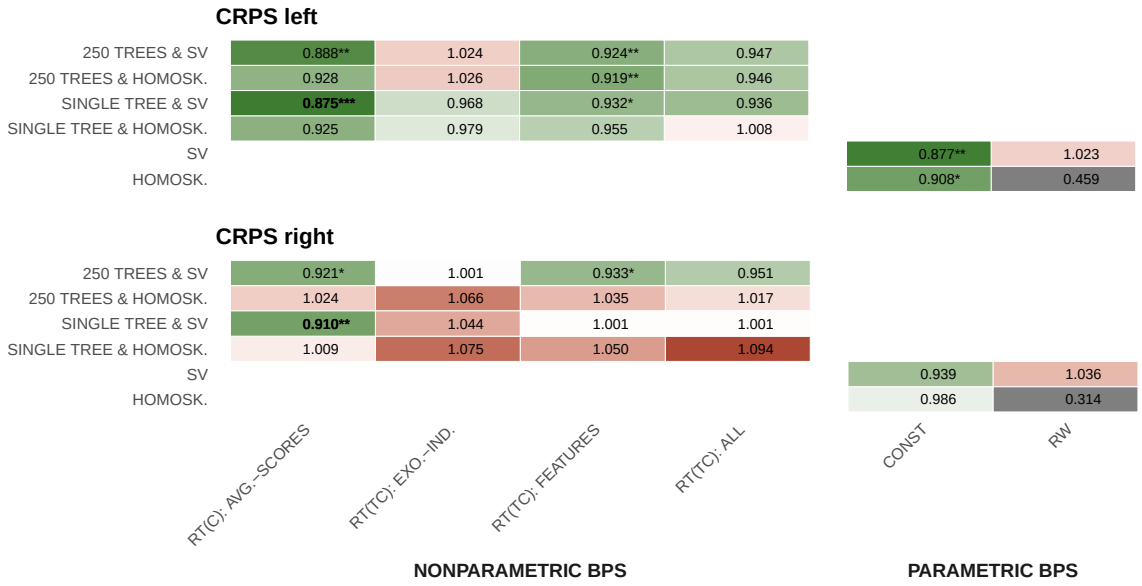
To estimate these ADL specifications, we use standard Bayesian non-conjugate regression techniques with posteriors of standard form. The non-conjugate priors are weakly informative. We center both ρ_π and α_π in Eq. (12) on a prior mean of zero and assume a prior variance of 100. For the case of homoskedastic errors, we assume an inverse Gamma prior on $\sigma^2_{\pi,t+h} := \sigma^2_\pi \sim i\mathcal{G}(0.01, 0.01)$, while for SV we essentially use the setup sketched in Appendix A.1 (see Eq. A.2). However, non-conjugacy (and SV in particular) leads to predictive densities for which there is no closed-form solution. Therefore we use MCMC methods and predictive simulation.

B Empirical Appendix: Additional Results

This empirical appendix contains supplementary results as referenced in the main paper. It is structured as follows. Section B.1 reports left and right tail CRPSs. Section B.2 presents the combination weights. Section B.3 presents the cumulative CRPS statistics. Section B.4 presents the fluctuation tests. Section B.5 presents the PITs tests. Section B.6 presents results showing how we can draw out the degree of shrinkage implied by BPS-RT. Section B.7 plots the predictive densities in both applications and examines their skewness.

B.1 Tail forecast accuracy

Figure B.1: Relative tail forecast accuracy: EA GDP growth.



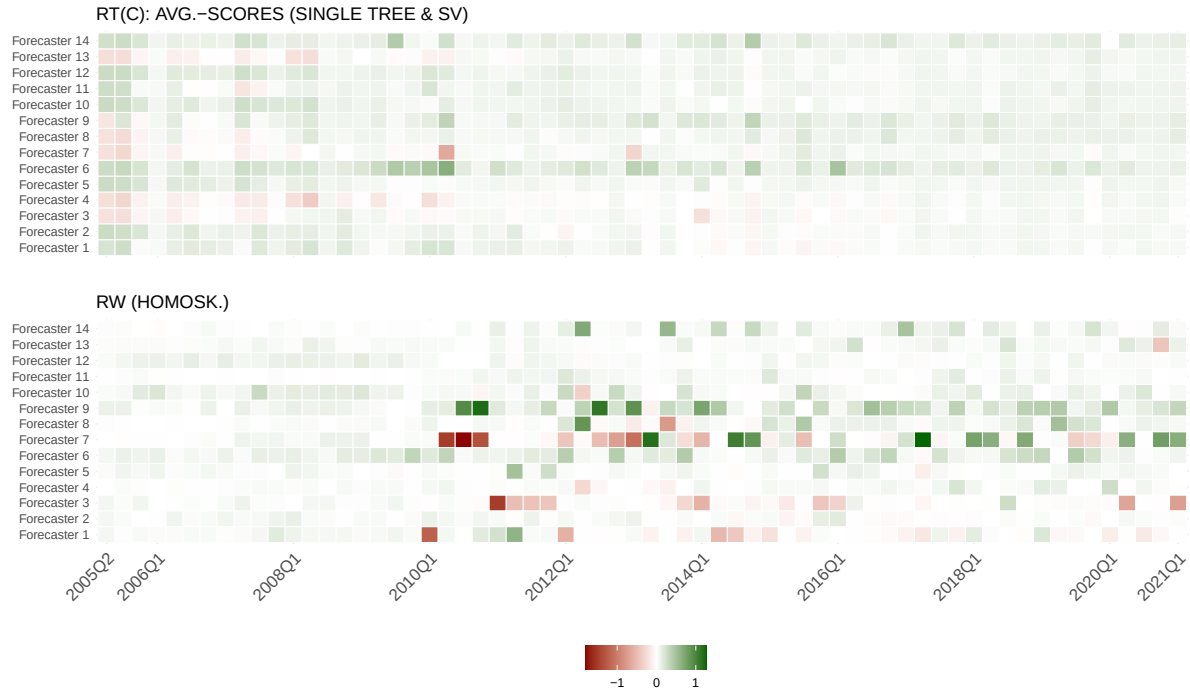
Notes: This figure shows two variants of quantile-weighted CRPS ratios, one focusing on the left tail and the other on the right tail. The gray-shaded entries give the actual scores of our benchmark (BPS-RW with homoskedastic error variances). Green-shaded entries refer to models that outperform the benchmark (with the forecast metric ratios below one), while red-shaded entries denote models that are outperformed by the benchmark (with the forecast metric ratios greater than one). The best-performing model specification by forecast metric is given in bold. Asterisks indicate statistical significance of the Diebold and Mariano (1995) test, which assumes equal forecast performance for each model relative to the benchmark, at the 1 (***), 5 (**), and 10 (*) percent significance levels.

Figure B.2: Relative tail forecast accuracy: US inflation.

Notes: This figure shows two variants of quantile-weighted CRPS ratios, one focusing on the left tail and the other on the right tail. The gray-shaded entries give the actual scores of our benchmark (BPS-RW with homoskedastic error variances). Green-shaded entries refer to models that outperform the benchmark (with the forecast metric ratios below one), while red-shaded entries denote models that are outperformed by the benchmark (with the forecast metric ratios greater than one). The best-performing model specification by forecast metric is given in bold. Asterisks indicate statistical significance of the Diebold and Mariano (1995) test, which assumes equal forecast performance for each model relative to the benchmark, at the 1 (***), 5 (**), and 10 (*) percent significance levels.

B.2 Combination Weights

Figure B.3: Combination weights over the evaluation sample: EA GDP growth



Notes: This figure shows the posterior median of the combination weights, $(\gamma_j + \beta_{jt+h})$, for each of the 14 SPF forecasters. Green (red)-shaded cells indicate that calibration parameters are above (below) zero. The top panel corresponds to our preferred BPS-RT specification, while the bottom panel corresponds to the benchmark.

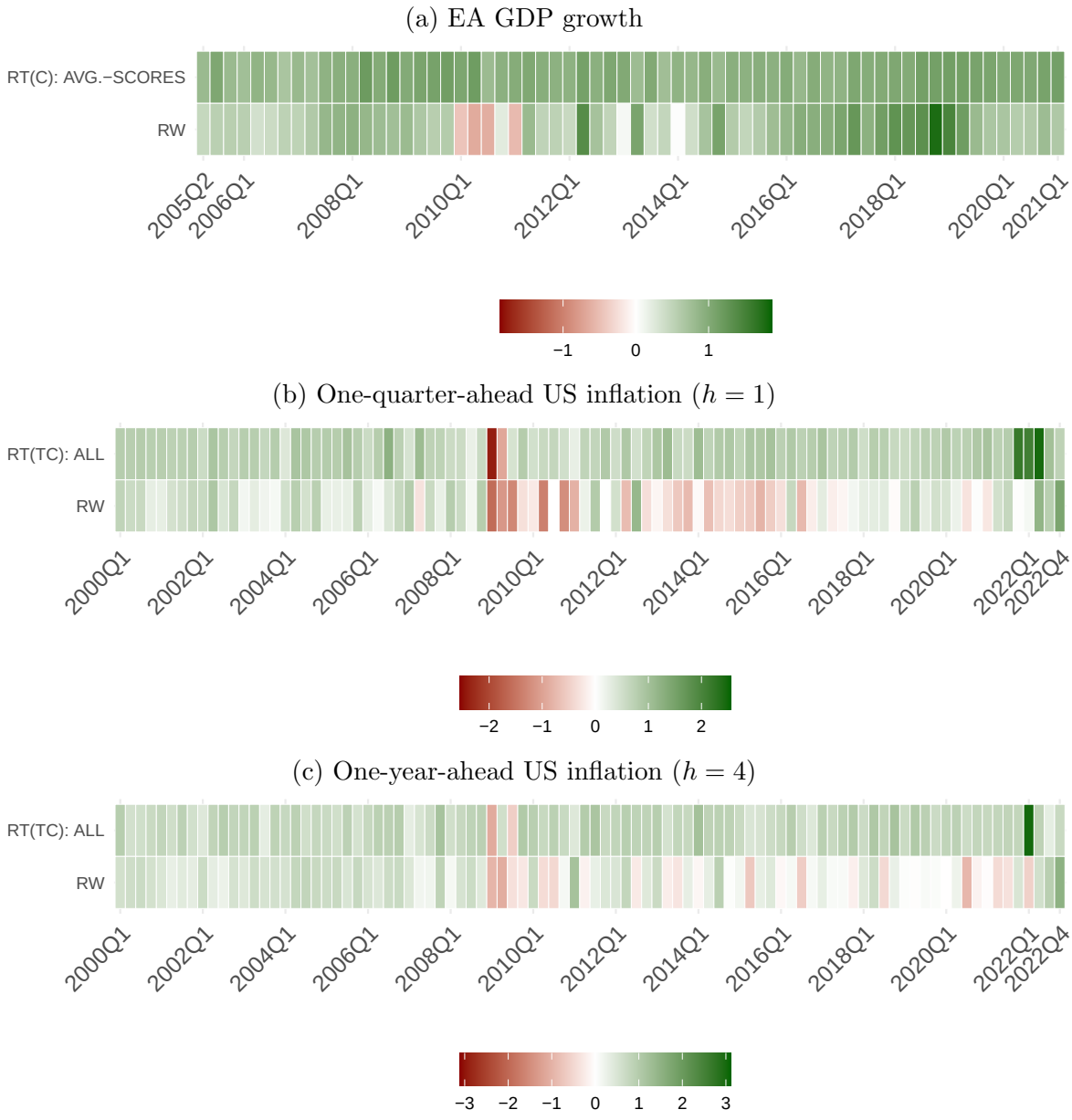
Figure B.4: Combination weights over the evaluation sample: One-quarter-ahead US inflation

Notes: This figure shows the posterior median of the one-quarter-ahead combination weights, $(\gamma_j + \beta_{jt+h})$, for each of the 56 ADL model variants. Green (red)-shaded cells indicate that weights are above (below) zero. The top panel corresponds to our preferred BPS-RT specification, while the bottom panel corresponds to the benchmark.

Figure B.5: Combination weights over the evaluation sample: One-year-ahead US inflation ($h = 4$)

Notes: This figure shows the posterior median of the one-year-ahead combination weights, $(\gamma_j + \beta_{jt+h})$, of the best-performing model parameters for each of the 56 ADL model variants. Green (red)-shaded cells indicate that weights are above (below) zero. The top panel corresponds to our preferred BPS-RT specification, while the bottom panel corresponds to the benchmark.

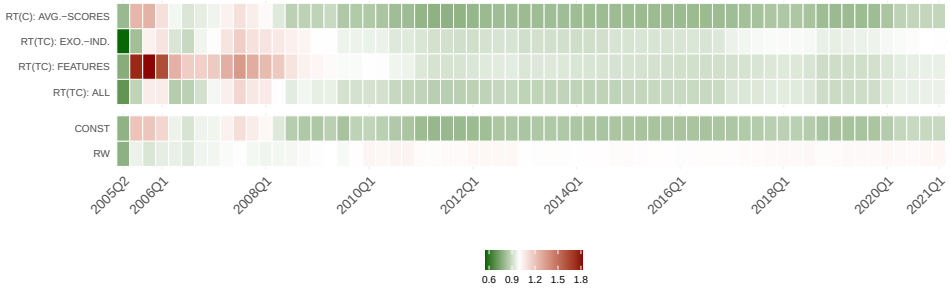
Figure B.6: Sum of combination weights over the evaluation sample



Notes: This figure shows the posterior median of the sum of the combination weights, $\sum_{j=1}^{J}(\gamma_j + \beta_{jt+h})$, for the models shown in Figures B.3, B.4 and B.5. Green (red)-shaded cells indicate that the overall sum of weights is above (below) zero for a specific evaluation period.

B.3 Cumulative CRPS

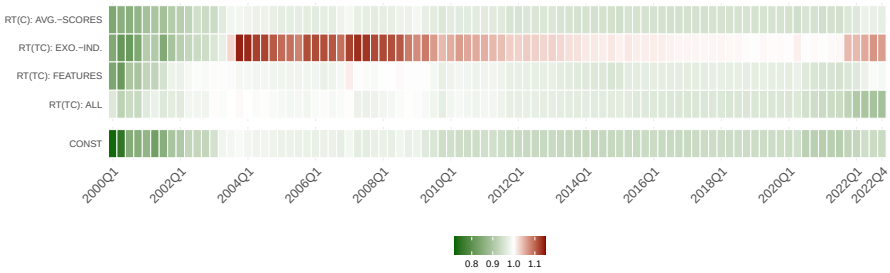
Figure B.7: Forecast performance of single-tree specifications with stochastic volatility: EA GDP growth



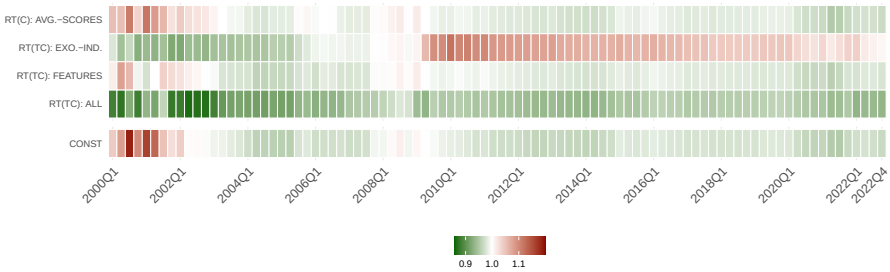
Notes: This figure shows relative cumulative continuous ranked probability scores (CRPSs) over the full evaluation sample, which ranges from 2005Q2 to 2021Q1. The benchmark model is a TVP regression with a random walk evolution of parameters (`BPS-RW`) and homoskedastic error variances. Green-shaded entries indicate periods in which the respective model outperforms the benchmark (with the cumulative CRPS ratio below one), while red-shaded entries denote periods in which the respective model is outperformed by the benchmark (with the cumulative CRPS ratio greater than one). We refrain from showing the forecast performance over time for all models, but focus on the class of models that contains the best-performing specification in terms of CRPS, that is, all single-tree specifications with stochastic volatility.

Figure B.8: Forecast performance of single-tree specifications with homoskedastic error variances: US inflation

(a) One-quarter-ahead ($h = 1$)



(b) One-year-ahead ($h = 4$)



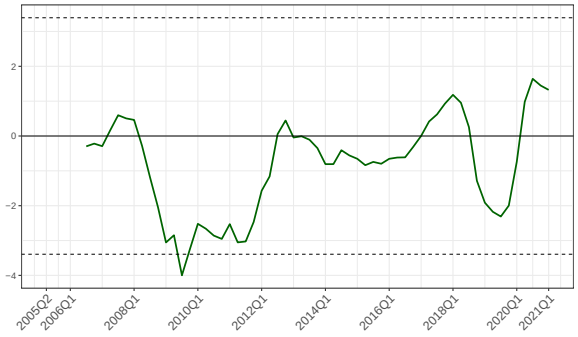
Notes: This figure shows average relative cumulative continuous ranked probability scores (CRPSs) over the evaluation sample, which ranges from 2000Q1 to 2022Q4. The benchmark model is a TVP regression with a random walk evolution of parameters (`BPS-RW`) and homoskedastic error variances. Green-shaded entries indicate periods in which the respective model outperforms the benchmark (with the cumulative CRPS ratio below one), while red-shaded entries denote periods in which the respective model is outperformed by the benchmark (with the cumulative CRPS ratio greater than one). We refrain from showing the forecast performance over time for all models, but focus on the class of models that forecast well in terms of CRPS, that is, all homoskedastic, single-tree specifications, the class containing the best-performing specification for the one-quarter-ahead ($h = 1$) horizon.

B.4 Giacomini and Rossi (2010) Fluctuation Test Statistic

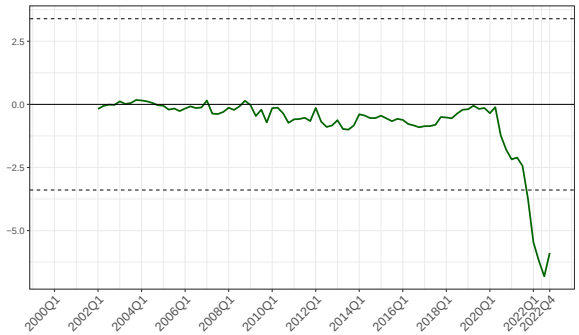
We focus on evaluating the BPS-RT specifications with the best overall forecast performance. As seen from Figures 2 and 3, in the EA-GDP application, this is the single-tree specification with SV using average scores as effect modifiers. For the US inflation application, this is the homoskedastic single-tree specification with the full set of weight modifiers.

Figure B.9: Evolution of the Giacomini and Rossi (2010) fluctuation test statistic

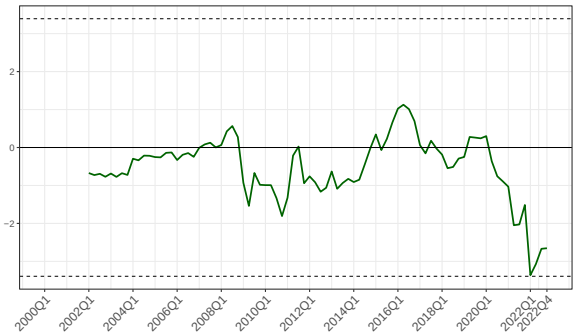
(a) EA GDP growth



(b) One-quarter-ahead US inflation ($h = 1$)



(c) One-year-ahead US inflation ($h = 4$)



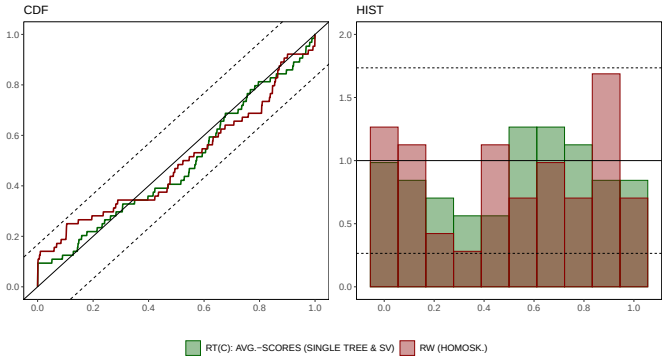
Notes: This figure shows the evolution of the Giacomini and Rossi (2010) fluctuation test statistic over time. The green solid line represents the test statistic, the black solid line marks the zero line, and the black dashed lines indicate the respective 95% confidence bands. To compute this period-specific test statistic, we use local relative continuous ranked probability scores (CRPSs) between the preferred BPS-RT specification and the benchmark (homoskedastic BPS-RW) over a rolling window comprising 10% of the evaluation sample. In panel (a), this implies that the rolling window is based on five observations (with the initial value of the test statistic available in 2006Q3), while in panels (b) and (c), this implies the rolling window is based on eight observations (with the initial value of the test statistic available in 2002Q1).

B.5 Probability Integral Transforms (PITs)

We focus on evaluating the BPS-RT specifications with the best overall forecast performance. As seen from Figures 2 and 3, in the EA-GDP application, this is the single tree specification with SV using average scores as effect modifiers. For the US inflation application, this is the homoskedastic single tree specification with the full set of weight modifiers.

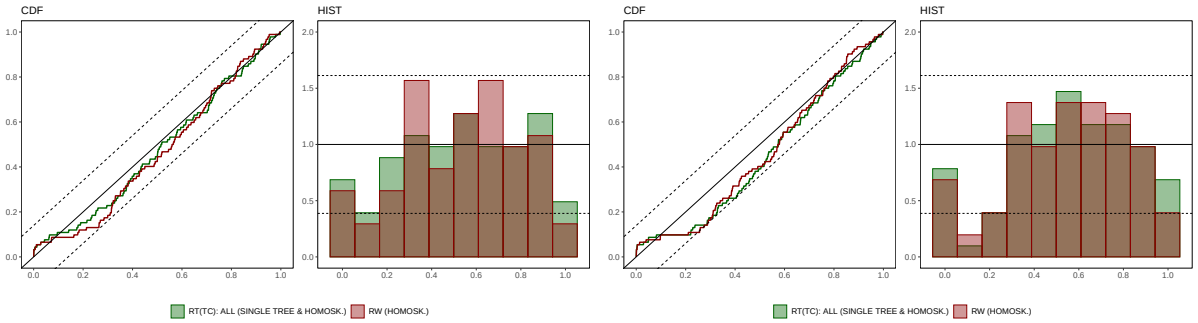
Figure B.10: Evaluating model calibration using probability integral transforms (PITs)

(a) EA GDP growth



(b) One-quarter-ahead US inflation ($h = 1$)

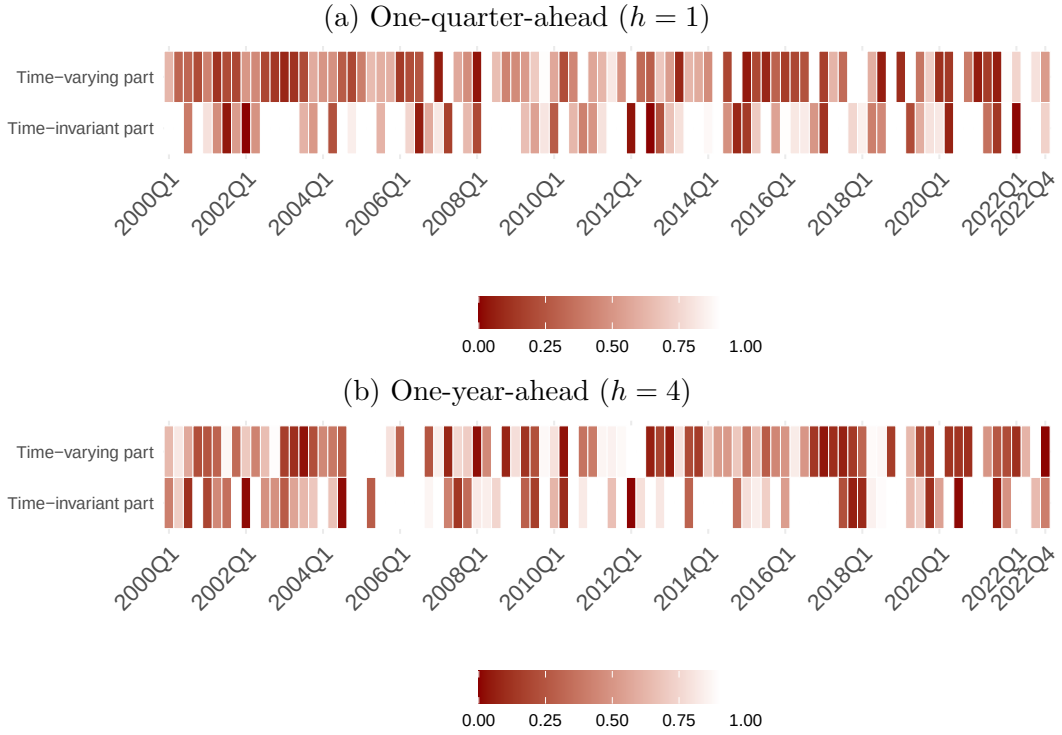
(c) One-year-ahead US inflation ($h = 4$)



Notes: This figure shows the empirical cumulative density function of the PITs in the left panels and the histogram of the PITs in the right panels. A correctly specified model has PITs that are standard uniformly distributed. Such a specification is denoted by the black solid lines, with the black dashed lines denoting the respective 95% confidence bands (see Rossi and Sekhposyan, 2019). The preferred BPS-RT specification is shown in green, while the benchmark is indicated in red.

B.6 Measuring the Degree of Shrinkage

Figure B.11: Overall degree of shrinkage toward the prior mean for US inflation.



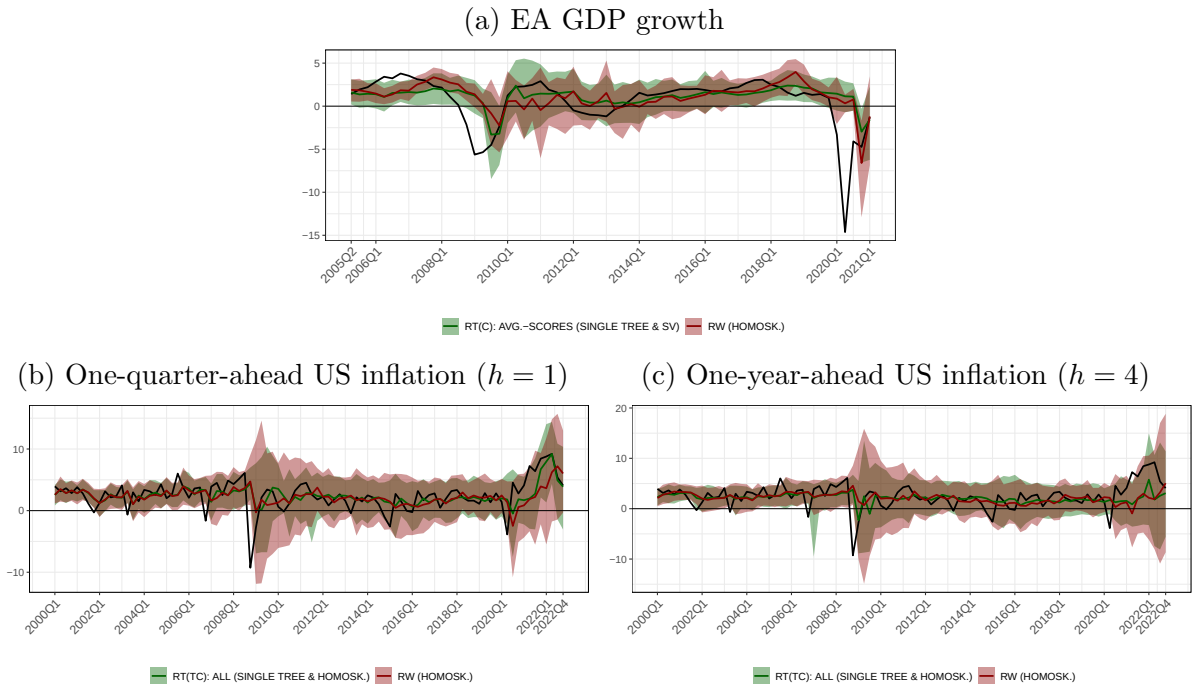
Notes: This figure shows the evolution of the degree of shrinkage measure over time. For each period in the evaluation sample, this measure is computed for our preferred specification (homoskedastic `BPS-RT(TC): ALL` with a single tree) as the ratio between the variation explained of the prior mean and the total variation of the respective coefficient part. This measure is bounded between zero and one. Values close to zero suggest that idiosyncratic deviations of coefficients (via the state innovation variances) dominate the prior mean (at least for some coefficients), while values close to one indicate that all coefficients are strongly pushed toward the prior mean in the respective period in the evaluation sample.

Figure B.11 shows a measure for overall shrinkage for both the time-invariant part ($\boldsymbol{\gamma}$) and the time-varying part ($\boldsymbol{\beta}_t$) of the weights. This measure effectively summarizes the overall variation explained by the prior mean as part of the overall variation in coefficients. It thus allows us to assess the relative importance of idiosyncratic innovations to the coefficients (i.e., innovations to the state equation) compared to the prior mean, and thus serves to quantify the overall degree of shrinkage by resembling something like a "joint" R^2-type of measure, which is bounded between zero and one. In each of the state equations, the target variables are either the constant coefficients $\boldsymbol{\gamma}$ or the stacked time-varying coefficients $\boldsymbol{\beta}_t$ (for, $t = 1, \dots, T$). In such a hierarchical model, the prior mean can then be treated as the conditional mean (i.e., the fit), while the state innovations (i.e., the shocks) are mainly driven by the state innovation (or prior) variances. In the following, a low joint R^2 suggests that the state innovation variances play a significant role (at least for some of the coefficients), whereas a high joint R^2 suggests that the

coefficients are heavily shrunk toward the prior mean. It is worth noting that in recessions, the R^2 is typically essentially zero and thus the prior means are less informative in these periods and random innovations to the state equations provide/add more model flexibility, which is required/necessary in these highly volatile periods.

B.7 Combined Predictive Densities and Their Skewness

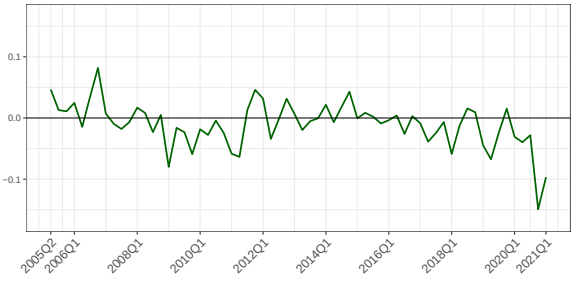
Figure B.12: BPS-RT and BPS-RW predictive densities



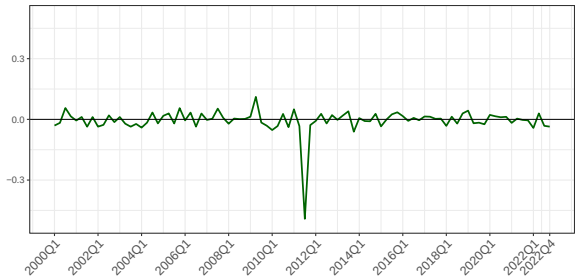
Notes: This figure displays the corresponding predictive densities. The colored shaded areas and the colored solid lines represent the 90% confidence interval and the posterior median, respectively. The preferred BPS-RT specification is shown in green, while the benchmark is indicated in red. The black solid line in both panels refers to the respective realization.

Figure B.13: Evolution of a quantile-based skewness measure for the BPS-RT predictive densities

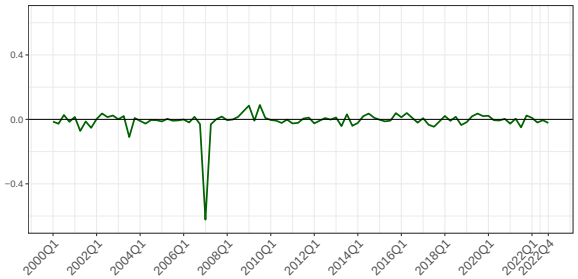
(a) EA GDP growth



(b) One-quarter-ahead US inflation ($h = 1$)



(c) One-year-ahead US inflation ($h = 4$)



Notes: This figure shows the evolution of a quantile-based skewness measure for predictive densities in our preferred BPS-RT specification, as shown in Figure B.12. The quantile-based skewness measure is defined as $((q_{95\%} - q_{50\%}) - (q_{50\%} - q_{5\%}))/(q_{95\%} - q_{5\%})$, where $q_{5\%}$, $q_{50\%}$, and $q_{95\%}$ represent the 5^{th}, 50^{th}, and 95^{th} percentiles of the predictive densities, respectively. The green solid line represents the computed skewness measure, while the black solid line marks the zero line.