INTERNATIONAL ECONOMIC REVIEW
Vol. 0, No. 0, February 2025 DOI: 10.1111/iere.12759

PREDICTIVE DENSITY COMBINATION USING BAYESIAN MACHINE LEARNING*

By Tony Chernis, Niko Hauzenberger, Florian Huber, Gary Koop, and James Mitchell

Bank of Canada, Canada; University of Strathclyde, U.K.; University of Salzburg, Austria; Federal Reserve Bank of Cleveland, U.S.A.

Based on agent opinion analysis theory, Bayesian predictive synthesis (BPS) is a framework for combining predictive distributions in the face of model uncertainty. In this article, we generalize existing parametric implementations of BPS by showing how to combine competing probabilistic forecasts using interpretable Bayesian tree-based machine learning methods. We demonstrate the advantages of our approach—in terms of improved forecast accuracy and interpretability—via 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 U.S. inflation produced by many simple regression models.

1. INTRODUCTION

Forecasts of macroeconomic and financial variables are used by forward-looking decision makers, not least by central banks given that changes in their policy tools take time to have an impact on the macroeconomy. This requires them to set monetary policy targeting forecasts of future macroeconomic values (e.g., see Woodford, 2007). Increasingly, as popularized by the Bank of England's "fan charts" for inflation and GDP growth, these forecasts take the form of density forecasts and thus provide a full probabilistic representation of the uncertainty inherent in any single-valued (point) forecast. The decision maker is then aware of the risks associated with the forecast and can extract probability event forecasts of specific interest and/or credible intervals from the underlying density forecast.

As well as the uncertainty associated with any given forecast, there is uncertainty about the "best" forecasting model. Competing density forecasts can come from different reduced-form or structural macroeconomic models estimated by Bayesian or frequentist methods, and/or be subjective and come from surveys. One way or another, all these multiple density forecasts are likely wrong or "incomplete," to use terminology from Geweke (2010) discussed further below. So the decision maker should not select one single forecast alone. Instead, building on the idea of combining point forecasts (Bates and Granger, 1969), a growing literature has found that combining multiple density forecasts can be effective both in improving accuracy and in

*Manuscript received July 2024; revised November 2024.

The views expressed herein 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 the editor, two anonymous referees, conference and seminar participants at MacroFor, Notre Dame, OeNB Freitagsseminar, Örebro, the SoFiE summer school, Surrey, including Christiane Baumeister, Drew Creal, Helmut Elsinger, Markus Knell, Malte Knüppel, Christine de Mol, Eoghan O'Neill, Michael Pfarrhofer, Luca Rossini, Anna Stelzer, Mattias Villani, and Mike West 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). Please address correspondence to: Niko Hauzenberger, Department of Economics, University of Strathclyde, 16 Richmond St, Glasgow G1 1XQ, U.K. E-mail: *niko.hauzenberger@strath.ac.uk*.

reflecting forecasting practice at some central banks.[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). A related literature has recently shown how combined density forecasts can also be used to average impulse responses (Ho et al., 2024) and how they can be integrated into decision analysis (Tallman and West, 2023; Chernis et al., 2024).

In this article, we extend existing density forecast combination strategies by proposing a more flexible and general nonparametric approach. This allows the competing density forecasts to be formally combined in a nonlinear and time-varying manner, with the combination weights on the competing density forecasts determined by information that may be external to the forecasting models. This information might include time-varying or time-invariant characteristics of the competing density forecasts not directly reflected in their predictive distributions. Or it might capture other common time-varying factors, such as general information that the decision maker may have about the macroeconomic environment or financial conditions, external to the forecasting models. Our approach reflects the reality that decision makers informally likely aggregate multiple density forecasts neither linearly nor independently of additional information they may have on current macroeconomic and financial conditions. We show how, despite their statistical flexibility and use of Bayesian machine learning algorithms, it is still possible to interpret the proposed combined density forecasts.

Key to our new model combination strategy is Bayesian predictive synthesis (BPS). BPS has emerged, as extended into a time-series context by McAlinn and West (2019), as a general method of combining density forecasts with a strong theoretical basis. BPS draws on an earlier Bayesian literature on agent or expert opinion analysis (West and Crosse, 1992). It provides a formal and theoretically justified method for pooling densities. BPS can be shown to nest many previous approaches, including the "linear opinion pool" widely used in the aforementioned literature (see, e.g., subsection 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 article, 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 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 on competing density forecasts to evolve smoothly over time. But such an assumption 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. A random walk process for the weights also provides little explanation for why individual models receive weight. It does not directly link temporal changes in the weights to external driving information.

These considerations motivate this article. 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, which can incorporate additional (external) information known to the decision maker, to specify the synthesis function. Specifically, we use regression trees to learn the law of motion governing the coefficients of a (conditionally) linear synthesis function. Accordingly, we label our version of BPS, BPS-RT, or simply, RT.

Regression tree methods are a natural way to let the model learn the law of motion governing the combination weights, as specified in the synthesis function. In conventional (single-

[1] See, among many others, Mitchell and Hall (2005), Wallis (2005), Hall and Mitchell (2007), Geweke and Amisano (2011), Koop and Korobilis (2012), Waggoner and Zha (2012), Billio et al. (2013), Aastveit et al. (2014), Conflitti et al. (2015), Kapetanios et al. (2015), Chernis and Webley (2022), Knotek and Zaman (2023), Aastveit et al. (2023), Čapek et al. (2023), and Diebold et al. (2023). Aastveit et al. (2019) survey and provide references to uses of density forecast combinations in fields beyond macroeconomics.

model) forecasting applications, tree-based models of the conditional mean have proven highly successful (see, e.g., Clark et al., 2023; Huber and Rossini, 2022; Huber et al., 2023). Crucially for our purposes, regression-tree methods require the choice of covariates, which we call "weight modifiers." These weight modifiers help determine the weights attached to the density forecasts. In contrast, conventional implementations of BPS model the combination weights as random walks and any relevant information in the form of additional (external) covariates is neglected.[2] For example, decision makers may wish to let the weights on 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 both "global" (i.e., macro information not associated with a particular forecaster) and "local" (i.e., micro information associated with a given forecaster) variables when determining the weights. The weights on each density forecast are dynamically determined via a sequence of decision rules. This way RT learns how to combine predictions in a highly flexible way using all relevant information contained in the predictive densities and the weight modifiers. We can then use the properties of the decision rules to tease-out which weight modifiers are most important in determining the combination weights. In order to help the decision maker further interpret the combined density and better understand the role each individual density is playing in the combination, we use a conditionally linear synthesis function. We show how this means that RT can be used to understand the role of model incompleteness, agent clustering, and the time-varying importance of the different weight modifiers.

Our article is related to the growing literature in macroeconomic modeling using machine learning and nonparametric methods to combine density forecasts. A small number of other papers have used nonparametric techniques to combine predictive densities (e.g., Bassetti et al., 2018, 2023; Jin et al., 2022). However, unlike our proposed method, these papers neither use regression trees nor fit explicitly within the formal BPS framework. Unlike popular approaches that use regression trees to estimate nonparametric regressions, we use them to model how the weights change over time as a function of additional covariates. Our approach is related to a recent paper by Farrell et al. (2020) that models the parameters of economic models as unknown functions of additional covariates, capturing heterogeneity in a flexible manner. Similarly, papers such as Creal and Kim (2021), Deshpande et al. (2024), Coulombe (2024), and Hauzenberger et al. (2023) treat the parameters of regression models nonparametrically instead of assuming a known functional relationship between the covariates and the endogenous variables.

The next section of the article introduces and motivates BPS in theory and then discusses how it has been implemented in the existing literature. It then proposes our generalization, RT, and explores its properties. Section 3 demonstrates the utility of 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 U.S. 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 RT produces well-calibrated and accurate forecasts. Notably, we find that single-tree models perform as well as or, in some cases, even better than multiple tree models. This finding stands 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 RT stems from its better ability to explain periods of volatility, such as the global financial crisis that affected euro area (EA) GDP growth and the post-COVID inflation period in the United States. Zooming in 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).

the best performing RT specification in the U.S. inflation application, we show how the combination forecasts from RT can be interpreted. 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 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 RT and Online Appendix B provides additional empirical results, as referenced in the main article.

2. BPS WITH REGRESSION TREES

In Subsections 2.1 and 2.2, we provide some background on BPS, distinguishing between BPS in theory and its use in practice in extant empirical applications. Then in Subsections 2.3 and 2.4, we explain how regression trees can be used to provide a more flexible way of operationalizing BPS.

2.1. *Bayesian Predictive Synthesis.* 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 actual 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 we assume that a decision maker is confronted with h-step-ahead forecast densities for variable $y_{\tau+h}$ for $\tau = 1 \ldots t$. These forecast densities are produced by J different agents, experts, or models. We label these predictive densities $\{\pi_{j\tau}(x_{j\tau+h|t})\}_{j=1}^{J}$. Thus, $x_{j\tau+h|t}$ is the forecast of $y_{\tau+h}$ made by agent j at time t and it is important to stress it is a random variable, not simply a point forecast. $\boldsymbol{x}_{t+h|t} = (x_{1t+h|t}, \ldots x_{Jt+h|t})'$ denotes the vector of forecast densities for all agents.

Agent opinion analysis theory (West, 1992; West and Crosse, 1992), extended to a time-series context by McAlinn and West (2019), shows that the optimal predictive density for y_{t+h} takes the form:

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

where $\alpha(y_{t+h}|\boldsymbol{x}_{t+h|t})$ denotes the synthesis function that reflects how the decision maker combines their prior information with the set of expert-based forecasts.[4]

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

[4] The synthesis function depends on unknown parameters and latent states that control the properties of the synthesis function but we have suppressed these to simplify the notation.

Theory offers no guide as to the specific choice of the synthesis function, $\alpha(y_{t+h}|\boldsymbol{x}_{t+h|t})$. 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 the synthesis function takes the form of a dynamic linear regression model with time-varying coefficients. Our synthesis functions also have a dynamic regression form. Similar to Chernis (2023), we use a noncentered parameterization (see Frühwirth-Schnatter and Wagner, 2010), which parameterizes the linear synthesis function in terms of a constant coefficient regression component, and a time-varying part, which captures deviations from constant coefficients:

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

where c_{t+h} is a time-varying intercept that, throughout the article, is 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.[5] The dynamics of $\boldsymbol{\beta}_{t+h}$ are a crucial modeling choice in BPS and the most common assumption is that the weights evolve according to independent random walks with innovation covariance matrix $\boldsymbol{V}$. This choice leads to a version of BPS that we label "RW." This model also assumes that the (log) variance $\varsigma_{t+h} := \log \sigma^2_{t+h}$ evolves 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.[6]

Note that, even though the synthesis function given in (2) assumes Gaussianity and linearity, it is very flexible. That is, although it is linear at a given point in time (which aids in interpretation), the weights in the linear combination can change over time. Furthermore, although (2) implies a Gaussian density conditional on all the parameters of the synthesis function (namely, $\boldsymbol{\beta}_{t+h}$, $\boldsymbol{x}_{t+h|t}$, c_{t+h}, and σ^2_{t+h}), when we carry out unconditional predictive inference, we marginalize out these parameters, leading to a predictive density that can be highly non-Gaussian. Additional flexibility is gained through the time-varying intercept and error variance. The reason we include them is related to the issue of model set incompleteness, discussed below.

2.2. *Properties of BPS.*

In this subsection, we briefly discuss some properties of BPS in general and then our particular parameterization of the synthesis function.

At a general level, and in contrast with some linear 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.[7] Moreover, the degree of temporal change in the weights of RW depends on the magnitude of the state innovation variances for these parameters: Small values imply slow, smooth adjustment of the weights over time; large values allow for bigger, sharper changes.

We emphasize two features of our parameterization of the synthesis function before we introduce our regression-tree approach, which provides a more flexible nonparametric rep-

[5] Outside of the BPS paradigm, we can also view (2) as a special case of density forecast combination strategies that take convolutions of densities from different models, where the weights are also allowed to vary flexibly both across time and models; see Billio et al. (2013) and Casarin et al. (2023).

[6] The prior choices for these parameters are given in Online Appendix A.1. Homoskedastic cases, which we also consider in our empirical work, 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.

[7] Linear density forecast combination methods often require the weights to lie between 0 and 1 and sum to 1. This facilitates Bayesian interpretation of the weights, as posterior probabilities in BMA, but even outside of the Bayesian framework the weights are often required to sum to unity to ensure that the combined density is a density (integrates to unity). However, the restriction that the weights be positive can be relaxed even in linear density forecast combinations; for discussion and references, see Genest and Zidek (1986). It is also of note that in linear point forecast combinations, as seen in Bates and Granger (1969) and Granger and Ramanathan (1984), there is also no need for the weights to be positive and/or to lie within the unit circle.

resentation of the synthesis function. First, as a special case, we can 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 that the combination weights are constant over time, "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 the decision maker'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 variance 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 subsection 2.2.3 of Tallman and West, 2023).[8] 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.

2.3. *Machine Learning the Synthesis Function.* In this article, our proposal is to relax the restrictions in 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 $\boldsymbol{\beta}_{t+h}$ and the constant component of these combination weights, $\boldsymbol{\gamma}$, in a nonparametric manner.

Conventional implementations of Bayesian machine learning methods, such as regression trees, cluster observations into groups. In our implementation, the "observations" are the combination weights attached to each agent's forecast density. The clustering is achieved based on some observed variables that we call "weight modifiers." These are denoted by a K_γ-vector $\boldsymbol{z}^\gamma_j$ (of constant weight modifiers) and a K_β-vector $\boldsymbol{z}^\beta_{jt+h|t}$ (of time-varying weight modifiers). These target $t+h$ but are known at time t and are application specific. In Subsection 3.1, we discuss their choice in more detail, but we note here that weight modifiers could be characteristics of the agent's forecasts (e.g., past forecast performance, forecast uncertainty, or higher moments) or indicators of macroeconomic conditions (e.g., uncertainty or financial conditions) such that the weights can change based on the state of the economy.

We postulate a general nonlinear relationship between the weights and the weight modifiers through unknown functions $\mu^\gamma_j(\boldsymbol{z}^\gamma_j)$ and $\mu^\beta_j(\boldsymbol{z}^\beta_{jt+h|t})$. In particular, we assume:

$$(3) \qquad \gamma_j \sim \mathcal{N}(\mu^\gamma_j(\boldsymbol{z}^\gamma_j), \tau^\gamma_j) \quad \text{and} \quad \beta_{jt+h} \sim \mathcal{N}(\mu^\beta_j(\boldsymbol{z}^\beta_{jt+h|t}), \tau^\beta_j),$$

where τ^γ_j and τ^β_j denote prior scaling parameters. For expositional convenience, we define $\mu^\gamma_j := \mu^\gamma_j(\boldsymbol{z}^\gamma_j)$ and $\mu^\beta_{jt+h} := \mu^\beta_j(\boldsymbol{z}^\beta_{jt+h|t})$. This flexible model differs from existing implementations of BPS through both the hierarchical priors used on elements in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{t+h}$ and by incorporating additional covariates into the decision maker's information set via the use of the weight modifiers.

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

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 Equation (3) into Equation (2) yields:

$$(4)\quad y_{t+h} = c_{t+h} + \sum_{j=1}^{J}\left[\underbrace{\left(\mu_j^{\gamma}(\boldsymbol{z}_j^{\gamma}) + \sqrt{\tau_j^{\gamma}}\nu_j^{\gamma}\right)}_{\gamma_j} x_{jt+h|t} + \underbrace{\left(\mu_j^{\beta}(\boldsymbol{z}_{jt+h|t}^{\beta}) + \sqrt{\tau_j^{\beta}}\nu_{jt+h}^{\beta}\right)}_{\beta_{jt+h}} x_{jt+h|t}\right] + \sigma_{t+h}u_{t+h},$$

with ν_j^{γ}, ν_{jt+h}^{β}, and $u_{t+h} \sim \mathcal{N}(0,1)$ 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 (4) introduces interaction terms. For instance, for the constant coefficient part of the synthesis function we have the form $\mu_j^{\gamma}(\boldsymbol{z}_j^{\gamma}) \times x_{jt+h|t}$. This specific form reduces the risk of overfitting by introducing structure on the space of functions that we approximate. Furthermore, since (4) is conditionally linear in $x_{jt+h|t}$, at any point in time interpretability is ensured, even though (4) is unconditionally nonlinear and time varying. We stress that we do not allow the individual forecasts themselves to be weight modifiers. That is, we do not let $\boldsymbol{z}_{jt+h|t}^{\beta}$ equal $x_{jt+h|t}$ (or $\boldsymbol{x}_{t+h|t}$). If there were equality, we would end up with a type of nonparametric regression model.[9]

This treatment can be contrasted with the alternative of treating the synthesis function, α, itself nonparametrically. As stated in the introduction, the main advantage of our approach is interpretability, since inspection of the elements of $\boldsymbol{\beta}_{t+h}$ lets us gauge the importance of a particular forecaster in shaping the combined forecast. Put differently, and using the words of Farrell et al. (2020), our approach models "heterogeneity" in the weights. Notice, however, that the dynamic nature of our model implies that heterogeneity can be either *static* (related to μ_j^{γ}) or *dynamic* (related to μ_j^{β}), or both.

2.4. *Implementing BPS with Regression Trees.* In principle, any technique can be used to infer the shape of the two functions, μ_j^{γ} and μ_j^{β}. For instance, neural networks (Farrell et al., 2020, 2021), Gaussian processes (Clark et al., 2024; Hauzenberger et al., 2025), or regression trees (Coulombe, 2024; Creal and Kim, 2021; Deshpande et al., 2024) might be used. Given its empirical success, we use Bayesian additive regression trees (BART; see Chipman et al., 2010) to estimate μ_j^{γ} and μ_j^{β} nonparametrically.

BART approximates the prior mean functions through a sum-of-trees model with S trees:

$$(5)\qquad \mu_j^{\gamma}(\boldsymbol{z}_j^{\gamma}) \approx \sum_{s=1}^{S} g(\boldsymbol{z}_j^{\gamma}|\mathcal{T}_s^{\gamma}, \boldsymbol{\phi}_s^{\gamma}) \quad \text{and} \quad \mu_{jt+h}^{\beta} = \mu_j^{\beta}(\boldsymbol{z}_{jt+h|t}^{\beta}) \approx \sum_{s=1}^{S} g(\boldsymbol{z}_{jt+h|t}^{\beta}|\mathcal{T}_s^{\beta}, \boldsymbol{\phi}_s^{\beta}),$$

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

[9] In this case, $\beta_{jt+h} = \mu_j^{\beta}(x_{jt+h|t}) + \sqrt{\tau_j^{\beta}}\nu_{jt+h}^{\beta}$ would be an unknown function of $x_{jt+h|t}$ and hence the forecasters could impact y_{t+h} in a nonlinear manner, jeopardizing interpretability. The only difference to a standard BART model is that this term would be additionally scaled by $x_{jt+h|t}$ in the synthesis function.

structures describe sequences of disjoint sets. These sets partition the input space (determined by $\boldsymbol{z}_j^{\gamma}$ 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 the vectors of weight modifiers, $\boldsymbol{z}_j^{\gamma}$ and $\boldsymbol{z}_{jt+h|t}^{\beta}$.

There are two main justifications for our RT modeling approach. First, there is no reason to restrict attention, as in RW, to random walk specifications for the evolution of β_{t+h}. 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 RW, an implicit assumption made is that the information set available to the decision maker comprises exclusively the agent-based forecast densities.[10] But, in principle, additional unmodeled information is available to the decision maker and might help inform evolution of the weights. In our RT approach, the weight modifiers, $\boldsymbol{z}_j^{\gamma}$ and $\boldsymbol{z}_{jt+h|t}^{\beta}$, comprise this extra information.

We estimate the tree structures and the terminal parameters alongside all other unknown parameters and therefore need to specify priors for them too. In doing so, we follow the recommendations of Chipman et al. (2010). One of the key advantages of BART is that little tuning of hyperparameters is necessary. Chipman et al. (2010) show that their recommended parameters work well across a wide variety of data sets and simulations. Similarly, several studies in economics have shown that these standard settings perform well (Clark et al., 2023; Hauzenberger et al., 2023; Huber and Rossini, 2022; Huber et al., 2023). An important modeling choice is the number of trees. We consider a range of values for S between 1 and 250. We begin with the $S = 1$ tree, motivated by the stylized fact that parsimony often works well when combining forecasts (see Timmermann, 2006). The much richer specification with $S = 250$ trees is motivated by the fact that the wider BART literature has found that adding in additional trees beyond 250 seldom improves performance.

The remaining model and prior specification choices are standard (follow the recommendations in the BART literature) and are discussed in detail in Online Appendix A. This technical appendix also describes the MCMC methods used to estimate RT. In summary, these MCMC methods use standard methods for predictive simulation from each individual model (to draw from each agent's forecast density) and then use standard methods for drawing from the regression-tree model conditional on the individual-agent draws. For RT, the algorithm for drawing the parameters of the regression trees is taken directly from Chipman et al. (2010).

2.5. *Pooling Properties of RT Combinations.*

RT uses regression trees as a prior that "pools" individual forecasts into groups based on the weight modifiers. In order to illustrate, note that the tree-based method of determining the weights is defined by disjoint sets that are determined by splitting rules. The rules are of the form $z_{k,jt+h|t}^{\beta} \leq d_k$ or $z_{k,jt+h|t}^{\beta} > d_k$, where $z_{k,jt+h|t}^{\beta}$ is the kth weight modifier for the jth agent/model and d_k is a threshold parameter associated with the kth weight modifier, which is estimated from the data. It is important to note, however, that any splitting rule associated with the kth weight modifier is common across agents and periods (i.e., 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_{jt+h}^{β}, this structure in a sense captures the notion of a pooling prior and reflects the situation that the decision maker decides on the weights associated with each of the different agents based

[10] 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 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.

on using additional factors $\boldsymbol{z}_j^\gamma$ and $\boldsymbol{z}_{jt+h|t}^\beta$ 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_j^γ 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 an RT model that assumes $\boldsymbol{\beta}_{t+h} = \boldsymbol{0}_J$ and features only a time-invariant part $\boldsymbol{\gamma}$, for which the prior mean $\boldsymbol{\mu}^\gamma = (\mu_1^\gamma, \dots, \mu_J^\gamma)'$ is defined by a single tree ($S = 1$) and by a single effect modifier in z_j^γ (i.e., z_j^γ is a scalar with $K_\gamma = 1$). In this case, the prior on γ_j can be written as:

$$\gamma_j = g(z_j^\gamma | \mathcal{T}_s^\gamma, \boldsymbol{\phi}_s^\gamma) + \sqrt{\tau_j^\gamma}\, v_j, \quad v_j \sim \mathcal{N}(0, 1). \tag{6}$$

If we now compute the difference between γ_j and γ_m for distinct agents, $j \neq m$, and assume that z_j^γ and z_m^γ 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:

$$(\gamma_j - \gamma_m) \sim \mathcal{N}(0, \tau_j^\gamma + \tau_m^\gamma). \tag{7}$$

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 RT.

Another feature of our prior is that the decision maker adjusts their 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.

2.6. *Illustrating RT Combinations.*

Our RT model has many bells and whistles. However, the main mechanisms can be explained easily using a simple toy example that turns off many of the features described above. This example is neither meant to be realistic nor to provide simulation evidence that our model works in a controlled environment. Rather, it is intended to illustrate how our approach allocates combination weights.

Assume that, unknown to the decision maker, the "true" data for y_t are generated by the following threshold model:

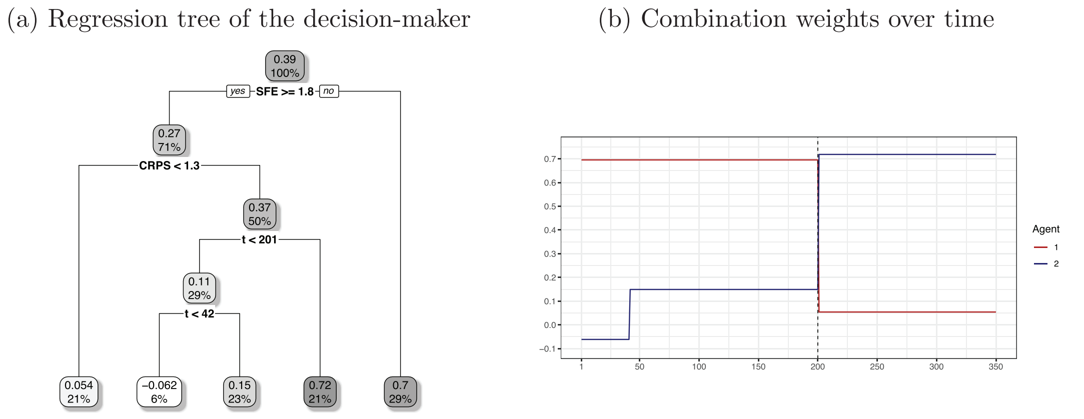
$$y_t = \begin{cases} \rho_1 y_{t-1} + c\rho_2 y_{t-2} + \sigma_0 \nu_t, & \text{for } t = 1, \dots, 200, \\ c\rho_1 y_{t-1} + \rho_2 y_{t-2} + \sigma_0 \nu_t, & \text{for } t = 201, \dots, 350, \end{cases} \tag{8}$$

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, $h = 1$):

$$x_{1t} \sim y_{1t} = \mathcal{N}(\rho_1 y_{t-1}, (1 - \rho_1^2)\sigma_0^2), \tag{9}$$

$$x_{2t} \sim y_{2t} = \mathcal{N}(\rho_2 y_{t-2}, (1 - \rho_2^2)\sigma_0^2). \tag{10}$$



NOTES: As weight modifiers, we use a simple linear time trend, as well as agent-specific squared forecast error (SFE) and continuous ranked probability score (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

ILLUSTRATION OF RT

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 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 RW is severely misspecified, since it 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. In order to capture historical point forecasting performance, we consider each agent's squared forecast error (SFE) as recursively computed at time t: $z^{\beta}_{2,jt+1|t} = (y_t - \mathbb{E}(x_{jt|t-1}))^2$ for $j = 1, 2$. Then to measure past density forecasting performance, we consider each agent's continuous ranked probability score (CRPS).[11]

Our synthesis function is given by Equation (2). In order to facilitate illustration of 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- and superscripts when there is no loss in meaning. Under these simplifying assumptions, the synthesis function, similarly to (4), reduces to:

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

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 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 $T = 350$.

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

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 a total of 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 the previous period 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% 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, the decision maker 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. In order 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, whereas 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, the decision maker now overweights the second agent, whereas the weight on Agent 1 is now much smaller.

This simple exercise illustrates how the decision maker incorporates additional information (time and past forecast errors in this case) to combine models. In general, though, the prior scaling parameters in RT are greater than zero, and hence, the regression 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 RT in two forecasting exercises. In the first application, we combine predictive densities of GDP growth for the EA produced by individual professional forecasters participating in the ECB SPF. Beyond its intrinsic interest, this data set is a good testing ground for RT because it has been used before when comparing alternative density forecast combination methods, including by Conflitti et al. (2015), Chernis (2023), and Diebold et al. (2023). Second, we forecast U.S. inflation using a set of autoregressive distributed lag (ADL) regression models. This data set and model set have been used by Stock and Watson (2003) and Rossi and Sekhposyan (2014), the latter using an ADL strategy similar to the one we deploy below 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 U.S. 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

Subsections 3.2 and 3.3. Both applications' evaluation samples cover the global financial crisis, the EA crisis, and the COVID-19 pandemic. Taken together, these two applications enable a comprehensive assessment of RT.

3.1. *Bayesian Machine Learning, Parametric, and Benchmark Specifications.* In order to draw out its properties, we consider, as set out in Subsection 3.1.1, a range of alternative specifications for RT, our Bayesian machine learning combinations. These are then compared against a set of competitor specifications as explained in Subsection 3.1.2. These comprise parametric BPS specifications and "simple" arithmetic and geometric density forecast combinations. With the aim of isolating the contribution of the individual forecasts to accuracy, we also compare both RT and parametric BPS combinations applied to the J individual forecasts, as in (1), with variants applied to aggregated forecasts. The rationale for this comparison is that, in (1), the BPS synthesis function, when estimating combination weights on the J individual forecasts, can both correct for potential miscalibration and capture any dependencies between the J density forecasts. By applying BPS to a forecast that is already combined, we strip out these effects. Specifically, we let the synthesis function take the form: $\alpha(y_{t+h}|\bar{x}_{t+h|t}) = \mathcal{N}\left(y_{t+h}|\bar{c}_{t+h} + \beta_{t+h}\bar{x}_{t+h|t}, \sigma^2_{t+h}\right)$, where the "average" forecast, $\bar{x}_t = (1/J)\sum_{j=1}^{J} x_{jt+h|t}$, is the arithmetic average. Variants of BPS that work off the average forecast, instead of exploiting all J individual forecasts, are labeled "BPS: AVG. AGENT."

The benchmark, against which we compare the forecast accuracy of every specification, is the traditional (parametric) implementation of BPS with time-varying weights (RW) and homoskedastic errors. Table 1 summarizes the full set of models and specifications used. We now explain each specification in turn.

3.1.1. *RT specifications.* We distinguish RT specifications according to the chosen weight modifiers, the number of trees, and whether RT is applied to the individual or combined forecast(s). Starting with the weight modifiers, we group and label them (cf. Table 1) as follows:

- **Cross-sectional ("local") weight modifiers** (z_j^{γ}): These allow the combination weights to vary by agent (but not over time, beyond the fact that even constant weight specifications will see the weights change through the out-of-sample window). In order to distinguish between "good" and "bad" forecasters, we score each agent's historical (ex post) forecast accuracy. Specifically, to capture historical point and density forecast accuracy, we set z_j^{γ} equal to agent-specific averages of SFE and CRPS. These averages are updated recursively through the out-of-sample window to reflect information only known in real time.
- **Time-varying ("global") weight modifiers** (z_{jt}^{β}): These indicators let the weights on agents vary over time to reflect economy-wide information. We consider a time trend (t), a measure of cross-sectional dispersion across agents defined as the standard deviation across the J experts' mean forecasts ("Disagr"), and the following application-specific measures designed to allow the combination weights to reflect the evolving state of the macroeconomy.
  - **EA SPF GDP application**: We use the European economic policy uncertainty (EPU) index of Baker et al. (2016).[12] In recessions, EPU rises; so allowing the combination weights to depend on it enables them to move with the business cycle.
  - **U.S. inflation application**: We consider households' one-year-ahead inflation expectations (IE) from the University of Michigan survey and the Chicago Fed's national financial conditions index (NFCI).[13] Both variables have been used to model inflation-at-risk (Lopez-Salido and Loria, 2024).
- **Cross-sectional and time-varying ("local") weight modifiers** (z_{jt}^{β}): These let the weights on individual experts change over time to reflect information local to both a given agent

[12] Available via https://www.policyuncertainty.com.

[13] These are available from the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org).

Table 1
overview of methods and specifications

Model Description			Weight Modifiers		
Class	Specification	TVP	Cross-Sectional	Time-Varying	Time & Cross-Sectional
			RT: Bayesian Machine Learning Combinations (Panel (a) in Figures 2 and 3)		
BPS: AVG. AGENT	EXO.-IND.	✓		Trend, EPU (EA), IE (US), NFCI (US)	
BPS: ALL AGENTS	SCORES		Avg. SFE, Avg. CRPS		
BPS: ALL AGENTS	TREND	✓		Trend	
BPS: ALL AGENTS	EXO.-IND.	✓	Avg. SFE, Avg. CRPS	Trend, EPU (EA), IE (US), NFCI (US)	
BPS: ALL AGENTS	FEATURES	✓	Avg. SFE, Avg. CRPS	Disagr	SFE, CRPS, Moments
BPS: ALL AGENTS	ALL	✓	Avg. SFE, Avg. CRPS	Disagr, Trend, EPU (EA), IE (US), NFCI (US)	SFE, CRPS, Moments
			Parametric Competitors and Simple Combinations (Panel (b) in Figures 2 and 3)		
AVG.	ARITHMETIC				
AVG.	GEOMETRIC				
BPS: AVG. AGENT	CONST				
BPS: AVG. AGENT	RW	✓			
BPS: ALL AGENTS	CONST				
			Benchmark Combination (gray shaded entries in Panel (b) in Figures 2 and 3)		
BPS: ALL AGENTS	RW	✓			

Note: This table summarizes properties of the six nonparametric Bayesian machine learning BPS-RT combinations, the five parametric BPS and simple competitors, and the benchmark combination model. “Class” indicates whether all J individual forecasts are combined using BPS (**BPS: ALL AGENTS**), whether BPS-RT is applied to forecasts already aggregated via an arithmetic average (**BPS: AVG. AGENT**), or whether a naïve average is considered (**AVG**.). For BPS-RT, “Specification” refers to the type of effect modifiers used. For parametric competitors, it specifies the law of motion (CONST vs. RW), whereas for **AVG**., it specifies the type of aggregation considered (ARITHMETIC vs. GEOMETRIC). “TVP” indicates with a tick if the model implies combination weights that are time-varying. Weight modifiers are divided into cross-sectional, time-varying, and time and cross-sectional. “Avg.” SFE and CRPS denote SFE and CRPS statistics averaged over the available sample, to contrast the use of time-varying SFE and CRPS statistics defined using agents’ forecasts made at time $(t-h)$. EPU is the economic policy uncertainty index. IE is the inflation expectations measure. NFCI is the Chicago Fed’s national financial conditions index. “Disagr” is the time t standard deviation of the J agents’ mean forecasts, and “moments” denotes use of the first four moments of each agent’s time t density forecast.

and a point in time, specifically their recent forecast performance and attributes of their latest forecast density. In order to capture recent forecast performance, we consider each agent's time t SFE and CRPS statistics, based on their forecasts made at time $(t - h)$, as available in real time. Then to allow the weights to cluster across agents, we also consider as elements of $\boldsymbol{z}_{jt}^{\beta}$ the first four moments of each agent's time t predictive density. This allows, among other things, for the possibility that high (ex ante) uncertainty forecasters should be weighted similarly.

For each set of weight modifiers, we implement BPS variants with homoskedasticity and with SV. In order to explore sensitivity to the number of trees, we consider BART specifications with $S = \{1, 50, 100, 250\}$ trees. Of note is the single tree ($S = 1$) specification that leads to a Bayesian regression-tree specification (see Chipman et al., 1998). 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 from $S = 1$, leads to an improvement in forecast performance. But this improvement tends to peter out as 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. But this may not be the case in BPS, since the data may prefer weights that are reasonably constant over time. As we shall see, we find that single-tree methods do indeed tend to forecast well in practice.

3.1.2. *Parametric and simple combinations.* We consider six competitor parametric combinations, with RW treated as the benchmark. Of the remaining five specifications, we examine BPS with constant weights, plus RW and constant weight BPS but applied to the average forecast (BPS: AVG. AGENT).

All these parametric BPS combinations make standard choices for the prior and MCMC method. Specifically, they are implemented in a fashion similar to that in Hauzenberger et al. (2022) and Chernis (2023). One change we make is to use the hyperparameter-free horseshoe prior, which shrinks parameter estimates toward zero. This way we have the same prior on the variance as with our regression-tree model. As the methods used are standard, we do not provide additional details here. Further econometric details are provided in Online Appendix A.

The remaining two competitor specifications are well-known "average" density forecasts. First, as already used as an input into "BPS: AVG. AGENT," we directly evaluate the arithmetic average of the J individual density forecasts. Use of such a "linear opinion pool" with equal weights, $1/J$, remains a common way of combining density forecasts. Second, we consider the geometric average, often called the logarithmic combination. This amounts to setting $\alpha(y_{t+h}|\boldsymbol{x}_{t+h|t}) = 1$ in (1).

Table 1 summarizes all the different specifications and the labels used to describe them when we present the forecasting results. As can be seen, we consider several permutations of the weight modifiers. We add to the set sequentially to assess the marginal benefit of each weight modifier and the benefit of including them all together. We remind the reader that for each specification in Table 1 we consider both homoskedastic and SV versions.

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.[14] We consider the two-quarters-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 instead of continuous density forecasts. For example, in the 1999Q1 survey, forecasters were instructed to

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

provide their probability forecasts over 10 bins. The first bin was GDP growth less than 0%, with the bins then increasing in intervals of 50 basis points, until the 10th bin of higher than 4% growth. In order to accommodate the discretized nature of these probability forecasts, instead of fitting 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 Online 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 nonresponders 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.[15]

We then take these 14 forecasters' densities and carry out a recursive out-of-sample evaluation 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 U.S. Inflation Using a Set of Indicators from FRED-QD.* We follow Rossi and Sekhposyan (2014) and construct density forecasts of U.S. inflation using a set of 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 United States. The selected indicators capture movements in asset 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 U.S. inflation to display significant heterogeneity. Table A.1 in the Online Appendix provides an overview of the variables used as exogenous predictors, their mnemonics, and the transformations applied to ensure their stationarity.

[15] 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.

We then use each of these ADL models to produce direct forecasts for quarter-on-quarter consumer price (CPIAUCSL) inflation one-quarter-ahead ($h = 1$).[16] Specifically, for each indicator, x_{jt}, for $j = 1, \ldots, 27$, we estimate the set of ADL models:

$$\pi_{t+h} = \rho_\pi \pi_t + \alpha_\pi x_{jt} + \varepsilon_{\pi,t+h}, \quad \varepsilon_{\pi,t+h} \sim \mathcal{N}\big(0, \sigma^2_{\pi,t+h}\big), \tag{11}$$

where π_t is inflation, ρ_π is the autoregressive coefficient, and α_π denotes the coefficient related to the jth exogenous indicator.[17] We supplement these $j = 1, \ldots, 27$ models with a 28th model (the AR(1) model) that sets $\alpha_\pi = 0$ in Equation (11). 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 Online 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 U.S. economy. The first 10 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 postpandemic inflationary period.

3.4. *Empirical Results.*

We break the empirical results into three parts presented in the following three subsections. First, we evaluate the relative and absolute density forecast accuracy of RT. Second, we examine why RT forecasts more accurately than other approaches by comparing features of their forecast densities. Third, we demonstrate aspects of interpretability of RT by examining how 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 instead of the point forecasts.[18] We measure the relative forecast accuracy of the forecast densities using two popular metrics: CRPS and a tail-weighted CRPS. The former evaluates the whole density, whereas the latter focuses on accuracy in the tails (Gneiting and Ranjan, 2011).[19] In the Online Appendix (see Subsection B.4), we also test the absolute calibration of the combined density forecasts using the Rossi and Sekhposyan (2019) test on the probability integral transforms. We show that the preferred RT specifications deliver well-calibrated density forecasts. We also indicate (in Subsection B.5 of the Online Appendix) how forecast performance fluctuates over time using the Giacomini and Rossi (2010) test.

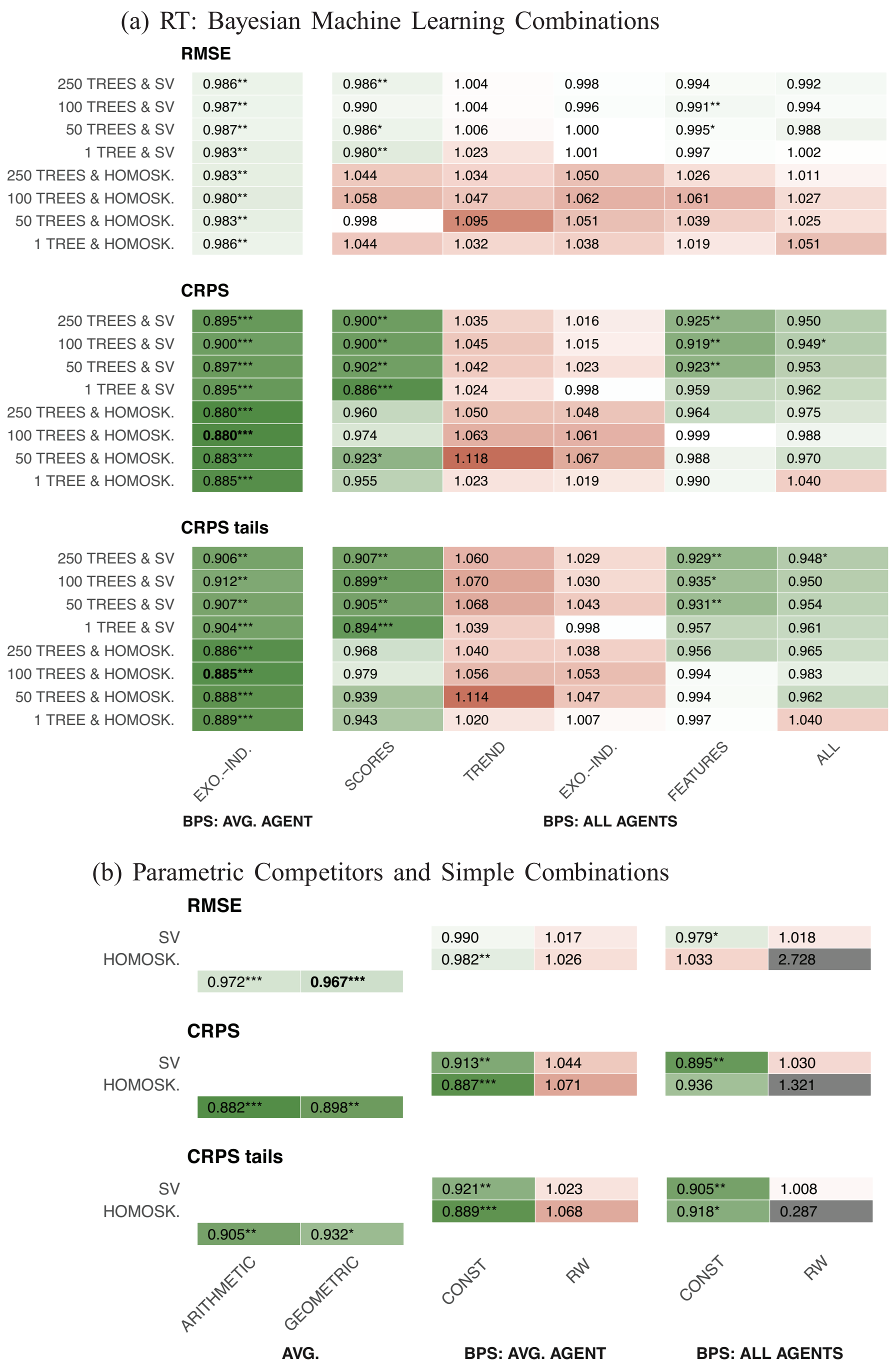
Figures 2 and 3 report the relative forecast performance of the different combinations in the EA GDP growth and U.S. inflation applications, respectively, using the RMSE, CRPS, and CRPS-tails loss functions. Each row in panel (a) of these figures reports the relative (to the RW benchmark) performance of the RT specifications as differentiated by the number of trees they use and whether they have SV or homoskedastic errors. The columns differentiate which set of weight modifiers is used and whether BPS is applied to all individual forecasts or

[16] We present the qualitatively similar one-year-ahead ($h = 4$) results in Figure B.1 in the Online Appendix.

[17] For notational ease, we do not use j subscripts to distinguish parameters in Equation (11).

[18] 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.

[19] Figures B.5, B.6, and B.7 in Online Appendix B follow Gneiting and Ranjan (2011) and break CRPS tails into their left and right tails.



(a) RT: Bayesian Machine Learning Combinations

RMSE

	EXO.-IND.	SCORES	TREND	EXO.-IND.	FEATURES	ALL
250 TREES & SV	0.986**	0.986**	1.004	0.998	0.994	0.992
100 TREES & SV	0.987**	0.990	1.004	0.996	0.991**	0.994
50 TREES & SV	0.987**	0.986*	1.006	1.000	0.995*	0.988
1 TREE & SV	0.983**	0.980**	1.023	1.001	0.997	1.002
250 TREES & HOMOSK.	0.983**	1.044	1.034	1.050	1.026	1.011
100 TREES & HOMOSK.	0.980**	1.058	1.047	1.062	1.061	1.027
50 TREES & HOMOSK.	0.983**	0.998	1.095	1.051	1.039	1.025
1 TREE & HOMOSK.	0.986**	1.044	1.032	1.038	1.019	1.051

CRPS

	EXO.-IND.	SCORES	TREND	EXO.-IND.	FEATURES	ALL
250 TREES & SV	0.895***	0.900**	1.035	1.016	0.925**	0.950
100 TREES & SV	0.900***	0.900**	1.045	1.015	0.919**	0.949*
50 TREES & SV	0.897***	0.902**	1.042	1.023	0.923**	0.953
1 TREE & SV	0.895***	0.886***	1.024	0.998	0.959	0.962
250 TREES & HOMOSK.	0.880***	0.960	1.050	1.048	0.964	0.975
100 TREES & HOMOSK.	**0.880*** **	0.974	1.063	1.061	0.999	0.988
50 TREES & HOMOSK.	0.883***	0.923*	1.118	1.067	0.988	0.970
1 TREE & HOMOSK.	0.885***	0.955	1.023	1.019	0.990	1.040

CRPS tails

	EXO.-IND.	SCORES	TREND	EXO.-IND.	FEATURES	ALL
250 TREES & SV	0.906**	0.907**	1.060	1.029	0.929**	0.948*
100 TREES & SV	0.912**	0.899**	1.070	1.030	0.935*	0.950
50 TREES & SV	0.907**	0.905**	1.068	1.043	0.931**	0.954
1 TREE & SV	0.904***	0.894***	1.039	0.998	0.957	0.961
250 TREES & HOMOSK.	0.886***	0.968	1.040	1.038	0.956	0.965
100 TREES & HOMOSK.	**0.885*** **	0.979	1.056	1.053	0.994	0.983
50 TREES & HOMOSK.	0.888***	0.939	1.114	1.047	0.994	0.962
1 TREE & HOMOSK.	0.889***	0.943	1.020	1.007	0.997	1.040

BPS: AVG. AGENT (EXO.-IND.) — BPS: ALL AGENTS (SCORES, TREND, EXO.-IND., FEATURES, ALL)

(b) Parametric Competitors and Simple Combinations

RMSE

	ARITHMETIC	GEOMETRIC	CONST	RW	CONST	RW
SV			0.990	1.017	0.979*	1.018
HOMOSK.			0.982**	1.026	1.033	2.728
	0.972***	**0.967*** **				

CRPS

	ARITHMETIC	GEOMETRIC	CONST	RW	CONST	RW
SV			0.913**	1.044	0.895**	1.030
HOMOSK.			0.887***	1.071	0.936	1.321
	0.882***	0.898**				

CRPS tails

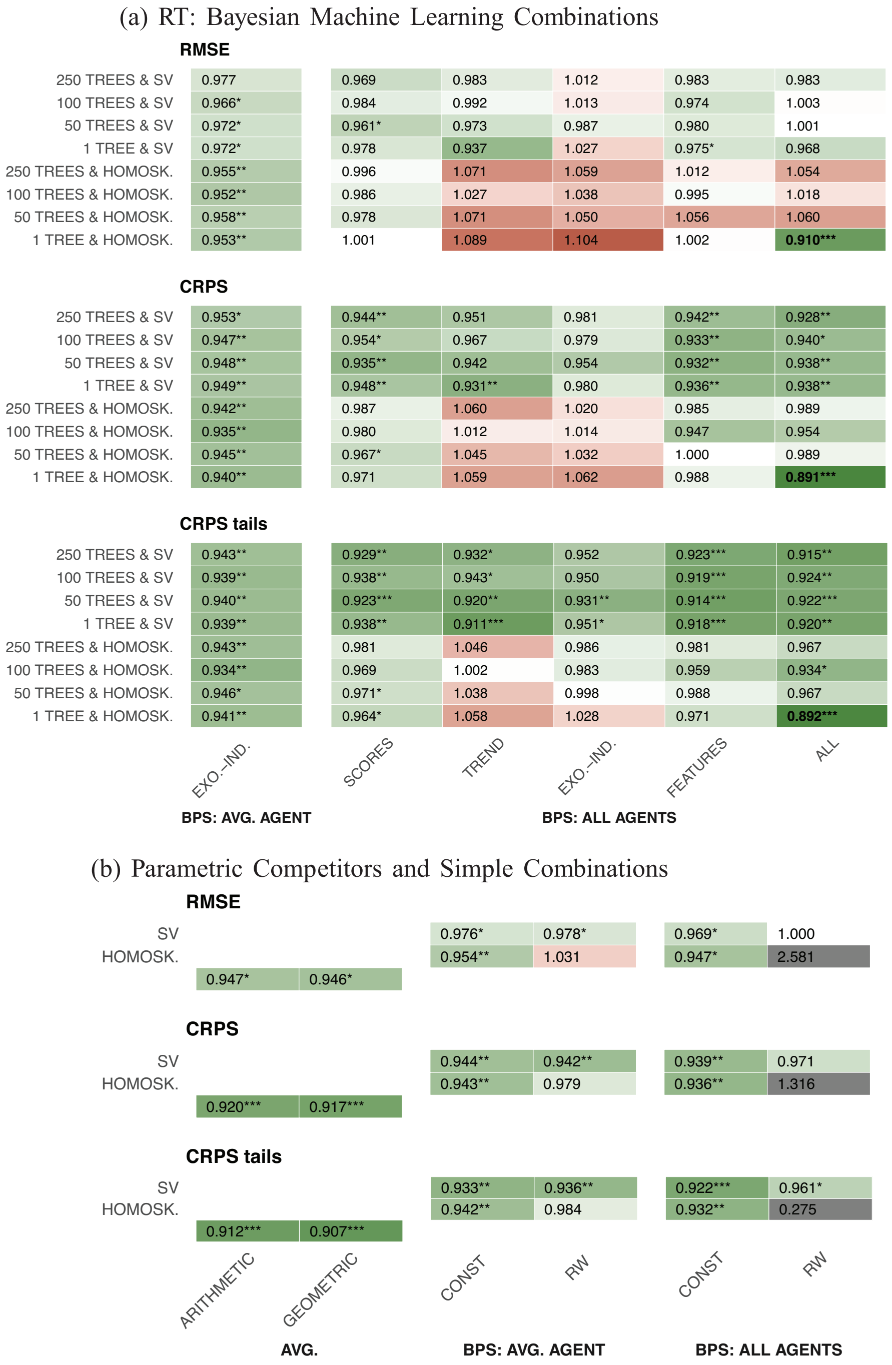
	ARITHMETIC	GEOMETRIC	CONST	RW	CONST	RW
SV			0.921**	1.023	0.905**	1.008
HOMOSK.			0.889***	1.068	0.918*	0.287
	0.905**	0.932*				

AVG. (ARITHMETIC, GEOMETRIC) — BPS: AVG. AGENT (CONST, RW) — BPS: ALL AGENTS (CONST, RW)

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. Gray shaded entries give the actual scores of our RW benchmark (BPS: ALL AGENTS (RW), with homoskedastic error variances). Green shaded entries refer to models that outperform the benchmark, whereas red shaded entries denote models that are outperformed by the benchmark. The best performing model specification by forecast metric is given in bold. Asterisks indicate statistical significance of the Diebold and Mariano (1995) test (testing each specification against the benchmark) at the 1% (***), 5% (**), and 10% (*) significance levels. See Table 1 for a summary of the different models.

FIGURE 2

RELATIVE FORECAST ACCURACY: EA GDP GROWTH



(a) RT: Bayesian Machine Learning Combinations

RMSE

	BPS: AVG. AGENT	BPS: ALL AGENTS				
	EXO.-IND.	SCORES	TREND	EXO.-IND.	FEATURES	ALL
250 TREES & SV	0.977	0.969	0.983	1.012	0.983	0.983
100 TREES & SV	0.966*	0.984	0.992	1.013	0.974	1.003
50 TREES & SV	0.972*	0.961*	0.973	0.987	0.980	1.001
1 TREE & SV	0.972*	0.978	0.937	1.027	0.975*	0.968
250 TREES & HOMOSK.	0.955**	0.996	1.071	1.059	1.012	1.054
100 TREES & HOMOSK.	0.952**	0.986	1.027	1.038	0.995	1.018
50 TREES & HOMOSK.	0.958**	0.978	1.071	1.050	1.056	1.060
1 TREE & HOMOSK.	0.953**	1.001	1.089	1.104	1.002	**0.910*** **

CRPS

	BPS: AVG. AGENT	BPS: ALL AGENTS				
	EXO.-IND.	SCORES	TREND	EXO.-IND.	FEATURES	ALL
250 TREES & SV	0.953*	0.944**	0.951	0.981	0.942**	0.928**
100 TREES & SV	0.947**	0.954*	0.967	0.979	0.933**	0.940*
50 TREES & SV	0.948**	0.935**	0.942	0.954	0.932**	0.938**
1 TREE & SV	0.949**	0.948**	0.931**	0.980	0.936**	0.938**
250 TREES & HOMOSK.	0.942**	0.987	1.060	1.020	0.985	0.989
100 TREES & HOMOSK.	0.935**	0.980	1.012	1.014	0.947	0.954
50 TREES & HOMOSK.	0.945**	0.967*	1.045	1.032	1.000	0.989
1 TREE & HOMOSK.	0.940**	0.971	1.059	1.062	0.988	**0.891*** **

CRPS tails

	BPS: AVG. AGENT	BPS: ALL AGENTS				
	EXO.-IND.	SCORES	TREND	EXO.-IND.	FEATURES	ALL
250 TREES & SV	0.943**	0.929**	0.932*	0.952	0.923***	0.915**
100 TREES & SV	0.939**	0.938**	0.943*	0.950	0.919***	0.924**
50 TREES & SV	0.940**	0.923***	0.920**	0.931**	0.914***	0.922***
1 TREE & SV	0.939**	0.938**	0.911***	0.951*	0.918***	0.920**
250 TREES & HOMOSK.	0.943**	0.981	1.046	0.986	0.981	0.967
100 TREES & HOMOSK.	0.934**	0.969	1.002	0.983	0.959	0.934*
50 TREES & HOMOSK.	0.946*	0.971*	1.038	0.998	0.988	0.967
1 TREE & HOMOSK.	0.941**	0.964*	1.058	1.028	0.971	**0.892*** **

(b) Parametric Competitors and Simple Combinations

RMSE

	AVG.		BPS: AVG. AGENT		BPS: ALL AGENTS	
	ARITHMETIC	GEOMETRIC	CONST	RW	CONST	RW
SV			0.976*	0.978*	0.969*	1.000
HOMOSK.			0.954**	1.031	0.947*	2.581
	0.947*	0.946*				

CRPS

	AVG.		BPS: AVG. AGENT		BPS: ALL AGENTS	
	ARITHMETIC	GEOMETRIC	CONST	RW	CONST	RW
SV			0.944**	0.942**	0.939**	0.971
HOMOSK.			0.943**	0.979	0.936**	1.316
	0.920***	0.917***				

CRPS tails

	AVG.		BPS: AVG. AGENT		BPS: ALL AGENTS	
	ARITHMETIC	GEOMETRIC	CONST	RW	CONST	RW
SV			0.933**	0.936**	0.922***	0.961*
HOMOSK.			0.942**	0.984	0.932**	0.275
	0.912***	0.907***				

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. Gray shaded entries give the actual scores of our RW benchmark (BPS: ALL AGENTS (RW), with homoskedastic error variances). Green shaded entries refer to models that outperform the benchmark, whereas red shaded entries denote models that are outperformed by the benchmark. The best performing model specification by forecast metric is given in bold. Asterisks indicate statistical significance of the Diebold and Mariano (1995) test (testing each specification against the benchmark) at the 1% (***), 5% (**), and 10% (*) significance levels. See Table 1 for a summary of the different models.

Figure 3

Relative forecast accuracy: U.S. inflation just the average forecast. Panel (b) shows analogous results for the parametric alternatives to RT, including the arithmetic and geometric averages.

Looking first at the RMSE results in panel (a) of Figure 2 for EA GDP growth, we see little difference between the alternative RT specifications in terms of their point forecast accuracy when using the individual forecasts (BPS: ALL AGENTS). Their accuracy is also similar to that of the benchmark (RW), with RMSE ratios around unity. Although the gains are still modest, albeit statistically significant, RT is consistently more competitive, irrespective of the number of trees or the presence of SV, when applied to the combined forecast (BPS: AVG. AGENT). This is supportive of the stylized fact from the forecasting literature that simple combination strategies, such as equal-weighted combinations of point forecasts, can be hard to beat (see Timmermann, 2006), especially when the individual forecasts all perform relatively similarly to each other. Comparing to the competitor forecasts (see panel (b) of Figure 2), we find that the geometric average is the single best performing combination strategy in terms of RMSE.

Turning to U.S. inflation (Figure 3), we see in the RMSE panel (panel (a)) that some of the tree-based methods more consistently improve upon the point forecast accuracy of the benchmark and in a manner that is statistically significant. Of particular note is the superior performance of the single-tree combinations, which frequently outperform models with more trees. We discuss this finding further below. Contrasting EA GDP, we also now see gains to combining the individual forecasts, instead of applying BPS to the average forecast or producing the arithmetic or geometric average.

The CRPS panels in Figures 2 and 3 reveal yet more of a payoff to using RT specifications, relative to RW, when we evaluate the whole density. Many of the forecast accuracy gains for RT are statistically significant. An implication of this finding is that RW's assumption that the combination weights follow a random walk is not supported by the data. But the CONST specifications remain competitive for EA GDP growth. We also find, in the inflation application, that RT can now produce more accurate forecasts when it exploits the individual forecasts.

The CRPS and CRPS tail results for inflation echo those under RMSE loss in concluding that using fewer tree structures tends to be preferable. The best performing RT combinations tend to have $S = 1$. 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, instead of the observed outcomes, nonparametrically, and hence, the implied conditional mean relation (see Equation (4)) introduces more restrictions relative to a standard BART model and hence lessens the risk of overfitting.

For density forecasts of U.S. inflation, we also find that adding SV improves forecasts. This is consistent with the findings when forecasting with (single) Bayesian models (see Clark, 2011). Looking at the CRPS and CRPS tails results for inflation (in Figure 3, panel (a)), we observe improvements from adding SV relative to the analogous model without SV. This is despite the fact that—and we touch on this again below when showing that these models in fact receive higher combination weights—in the U.S. inflation application, half of the components models themselves allow for SV. By contrast, the findings for EA GDP growth are somewhat more mixed. This is probably driven by the considerably smaller sample size.

Next we drill deeper into how specific choices of effect modifiers impact the performance of RT. This is accomplished by comparing forecast accuracy across the final 5 columns of panel (a) of Figures 2 and 3. This comparison reveals that the choice of weight modifier does affect RT's forecast accuracy. For the most part, just adding a time trend is not sufficient as there are additional gains in accuracy to letting the weights adapt to reflect agents' past forecast performance and/or features. But 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 RT specification) is consulted.

In summary, these results show that our machine learning combination schemes outperform existing parametric and simple density forecast combinations. But the gains are stronger in the U.S. inflation application. In the EA GDP application, given the smaller number of more similar agent-level density forecasts, it is, in our view, reassuring that RT, despite its greater flexibility, remains competitive with equal-weighted combinations. We know from previous research work that these simple combinations work well on this data set. In both applications, we see more substantial gains to RT when the density forecast, instead of just the point forecast, is evaluated. We also find significant performance variation across different RT specifications. For EA GDP growth, the best RT model employs 100 trees, exogenous indicators as weight modifiers, and homoskedastic error variances. But we emphasize again that single-tree RT variants perform extremely similarly. For U.S. inflation, the best-performing specification is a single-tree RT model using all the weight modifiers with homoskedastic shocks.

3.4.2. *Properties of the RT density forecasts.* In this subsection, we examine how and why RT forecasts more accurately. We focus on the best (most accurate) model in each application and compare its forecast densities to those of the benchmark model, RW.[20]

Figure 4 shows a heat map of the difference in probabilities, in intervals of 1.5% for EA GDP growth and of 1% for U.S. inflation, between RT and RW. Green (red) shading indicates that RT adds (subtracts) probability relative to 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.[21]

Panel (a) of Figure 4 shows that, in general, RT predictions are less dispersed than RW with more mass near the subsequent outcomes. Additionally, the RT density adds probability to low GDP growth outturns prior to the financial crisis and also forecasts higher growth than RW in both the post-global financial crisis recovery and the rebound from the COVID-19-induced recession.[22] Panel (b) of Figure 4 show the analogous plot for U.S. inflation. Similar to panel (a), RT places more mass closer to the outturn and produces forecasts that are, in general, less disperse. Moreover, RT adjusts much more quickly to the increase in inflation post-pandemic, attributing a higher probability to these outturns than RW. Consistent with the evidence in Rossi and Sekhposyan (2014) that combinations of predictive densities for U.S. inflation appear to be approximately Gaussian, the inflation forecast densities from RT also tend to be symmetric (see Figure B.15 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.

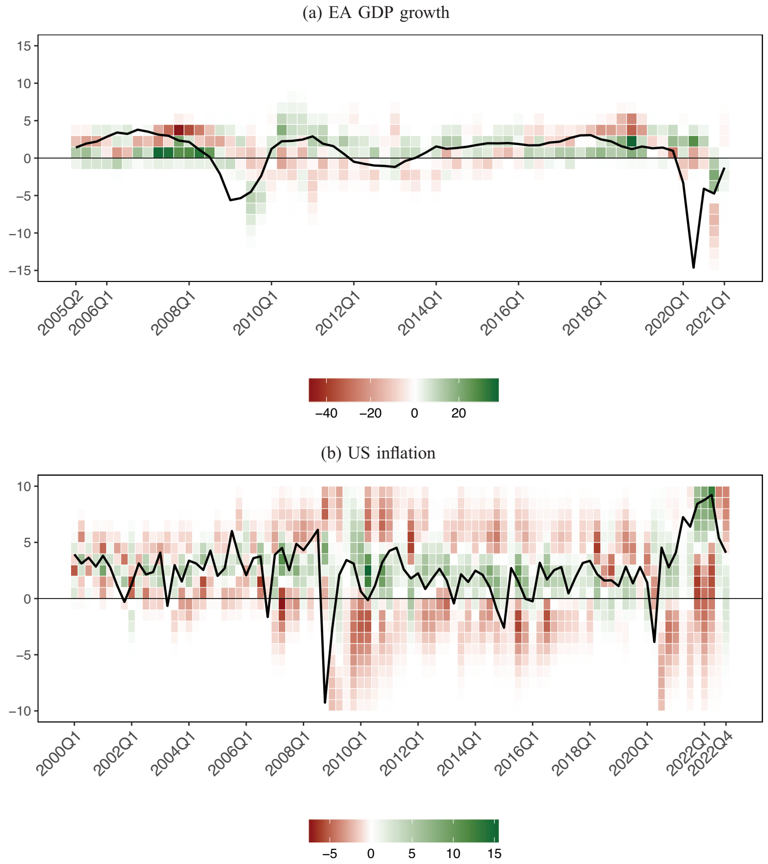
3.4.3. *Interpretation: A deeper dive into the mechanics of regression tree combinations for U.S. inflation.* This subsection explains how the decision maker can interpret the combined forecasts from RT. In so doing, we focus on the preferred RT specification in the U.S. inflation application, not least 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 RT is combining are actually able to forecast. Second, we assess the relative importance of individual weight modifiers in driving RT.

In order 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

[20] As seen in Figure 2 for the EA GDP growth application, the "best" RT specification—at least according to RMSE and it remains competitive for CRPS and CRPS tails too—has a single tree and SV and uses averages of SFE and CRPS as effect modifiers (i.e., SCORES). For the U.S. inflation application (see Figure 3), the "best" RT specification has a single tree, homoskedastic errors, and the full set of weight modifiers (i.e., ALL).

[21] For an alternative but complementary visualization, Figure B.14 in Online Appendix B shows the temporal evolution of the underlying density forecasts from RT and the benchmark RW model over the EA and U.S. evaluation samples.

[22] As shown in Figure B.15 in the Online Appendix, in moving the probability mass from the centers to the left tail of the forecast density, RT captures asymmetries in the forecast densities. Although 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.



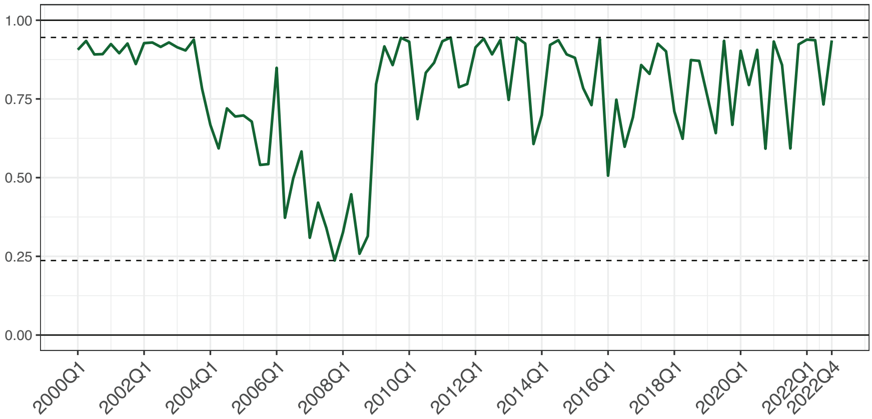
NOTES: This figure shows the difference in probabilities between the best performing RT model (in terms of CRPS) and RW. We define a grid of possible values for EA GDP growth ranging from −15% to 15% with increments of 1.5%, whereas we define a grid of possible values for U.S. inflation ranging from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

FIGURE 4

DIFFERENCE IN PROBABILITIES BETWEEN RT AND RW

is computed, for a specific period in the evaluation sample, as the ratio between the variation in the conditional mean in Equation (2) explained exclusively by the RT component—which is the conditional mean in Equation (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, Equation (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. We see increases in model incompleteness during the period 2004–08, 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



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 RT specification (homoskedastic BPS: ALL AGENTS (ALL), with a single tree) as the ratio between the variation explained exclusively by the 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 Equation (2), plays an important role, whereas values close to one indicate that the RT part explains most of the variation.

FIGURE 5

MEASURING MODEL INCOMPLETENESS: U.S. INFLATION

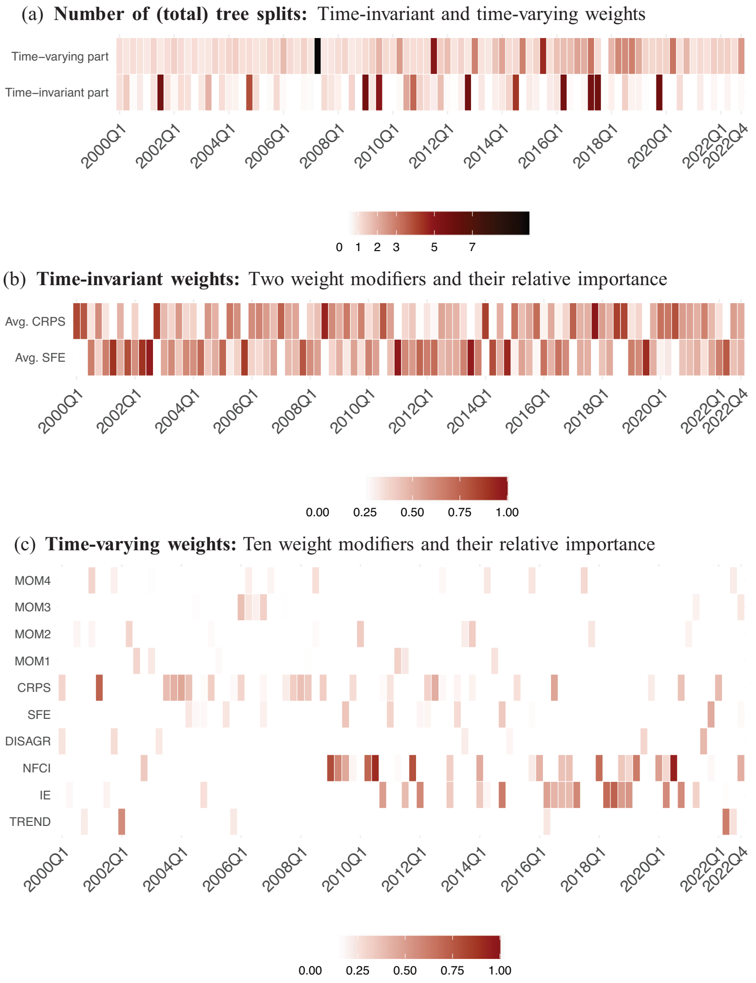
postpandemic rise in inflation, reinforcing the message from Figure 4 that RT was better able to anticipate the 2021 rise in U.S. inflation.

We now turn to assessing the relative importance of the individual weight modifiers in driving the density forecasts from 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. Figure 6 examines the weight modifiers forecasting U.S. inflation.[23]

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 RT. Panel (a) reveals that 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), whereas 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. Placing 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 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 often forecast well (see, e.g., 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 (average CRPS and average SFE) used to model the time-invariant combination

[23] Analogous results forecasting inflation one-year-ahead are reported in Figure B.4 in the Online Appendix and summarized below when the conclusions differ markedly from those discussed for the 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

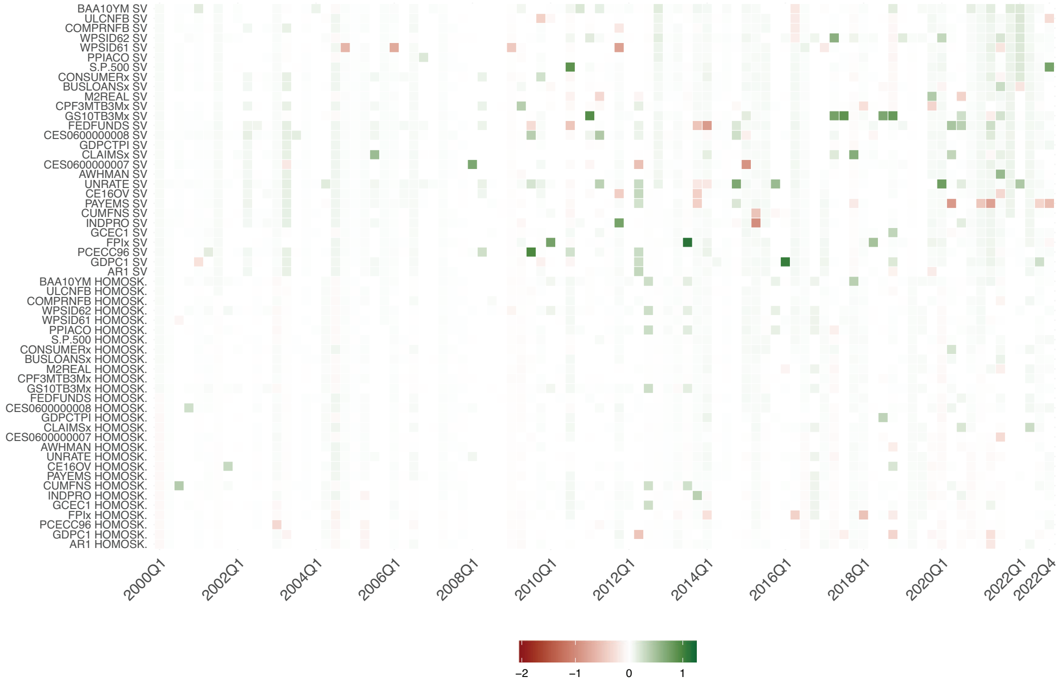
NOTES: Panel (a) shows the evolution of the total number of tree splits over time, whereas panels (b) and (c) show the marginal importance of each weight modifier for each quarter in the evaluation sample. Relative importance is defined as the share of the total number of splits. For each quarter in the evaluation sample, we obtain the posterior mean for these measures for our preferred RT specification (homoskedastic BPS: ALL AGENTS (ALL), with a single tree). MOM1 to MOM4 denote the first four moments of an agent's density. DISAGR is the standard deviation of the 56 agents' point forecasts made at time t, NFCI is the Chicago Fed's national financial conditions index, IE is the Michigan inflation expectations index, and TREND is a time trend.

FIGURE 6

NUMBER OF TREE SPLITS FOR RT $(S = 1)$ AND RELATIVE IMPORTANCE FOR EACH WEIGHT MODIFIER FOR U.S. INFLATION

weights. Neither CRPS nor SFE is obviously more important. Both weight modifiers receive positive and often fairly similar probabilities of inclusion. This implies that RT does partition models on the basis of their historical forecast accuracy.

Panel (c) of Figure 6 shows the importance of the time-varying weight modifiers. The first thing to notice is that there is much more sparsity in terms of the weight modifiers RT selects. For example, the various moments (MOM1-MOM4) and model dispersion (DISAGR) are relatively unimportant. Although 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 IE measure also receives more weight after the financial crisis. This is evidence that nonlinear



NOTES: This figure shows for our preferred RT specification the posterior median of the one-quarter-ahead combination weights, $(\gamma_j + \beta_{jt+h})$, on each of the 56 ADL agent forecasts. Green (red) shaded cells indicate that weights are above (below) zero. Table A.1 in the Online Appendix provides an overview of the variables used as exogenous predictors and their mnemonics.

FIGURE 7

COMBINATION WEIGHTS OVER THE EVALUATION SAMPLE: U.S. INFLATION

features of RT are driven by weight modifiers related to the business cycle. In other words, our 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, in Figure 7, we look at the posterior median estimates of the combination weights from the preferred RT specification over the evaluation sample to determine which agents get the most weight, that is, are the most useful forecasters. In this model-based inflation forecasting application, because the models only have one explanatory variable, we can thereby see which economic variables are most important. First, focusing on RT we see that the error variance specification is an important determinant of the weights. Models with stochastic volatility tend to receive larger weights. This is most apparent in the COVID-19 period. This corresponds to the period when RT outperforms the RW benchmark (see Figure B.12 in the Online Appendix). Additional results in the Online Appendix show that at the one-year-ahead horizon, the stochastic volatility models do also get more weight on average, but only slightly more than the homoskedastic models.

Second, there is a subset of agents (variables) that receives substantially more weight on average. These are models that are heteroskedastic and use either the unemployment rate (UNRATE), fixed private investment (FPIx), or the spread between 10-year government bonds and the three-month rate (GS10TB3Mx) as predictors. 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—those using GDP (GDPC1) and fixed investment (FPIx) with homoskedastic shocks and nonfarm payrolls (PAYEMS) with SV—get negative weights. This amounts to short-selling those models as a "hedge" against the models with correlated forecasts. Interestingly, this pattern is exaggerated for the one-year-ahead forecast combination weights where some real-activity variables get substantially negative weights (Figure B.9 in the Online Appendix).[24]

Although this subsection has focused on the U.S. inflation application, we end by returning briefly to the EA GDP growth forecasting application. Figure B.8 in the Online Appendix 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. That said, we do still see higher weights on a couple of agents (forecasters 6 and 14). We take this contrasting evidence across the two applications as empirical evidence that 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 article develops a method for nonparametric density forecast combination using regression trees, RT, and shows their value for probabilistic forecasting of macroeconomic variables in the face of model uncertainty. Although 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 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 test the performance of RT in two different applications—combining model-based U.S. inflation density forecasts and subjective histogram-based forecasts of EA GDP growth. We find that, across both applications, 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 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 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, 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

[24] Figure B.10 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 U.S. inflation this sum becomes negative during the global financial crisis, indicating how 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 RT can quickly adapt to temporal change.

structure. For instance, this could draw on the ability of 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 their own forecast (such as its mean or variance) but on characteristics of the other agents' forecasts.

DATA AVAILABILITY STATEMENT The data and codes that support the findings of this study are available at github.com/nhauzenb/chhkm-ier-bps. The data were obtained from publicly available resources. The ECB Survey of Professional Forecasters (SPF) data is accessible via https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/all_data.en.html, and the U.S. macroeconomic data can be obtained via https://www.stlouisfed.org/research/economists/mccracken/fred-databases.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table A.1: List of variables used for the autoregressive distributed lag (ADL) specifications.
Figure B.1: Relative forecast accuracy: US inflation, one-year-ahead (h = 4).
Figure B.2: Difference in probabilities between RT and RW, one-year-ahead US inflation (h = 4).
Figure B.3: Measuring model incompleteness: US inflation, one-year-ahead (h = 4).
Figure B.4: One-year-ahead horizon: Number of (total) tree splits for our single-tree models and relative importance for each effect modifier.
Figure B.5: Relative tail forecast accuracy: EA GDP growth.
Figure B.6: Relative tail forecast accuracy: US inflation, one-quarter-ahead (h = 1).
Figure B.7: Relative tail forecast accuracy: US inflation, one-year-ahead (h = 4).
Figure B.8: Combination weights over the evaluation sample: EA GDP growth
Figure B.9: Combination weights over the evaluation sample: One-year-ahead US inflation (h = 4)
Figure B.10: Sum of combination weights over the evaluation sample
Figure B.11: Evaluating model calibration using probability integral transforms (PITs)
Figure B.12: Evolution of the Giacomini and Rossi (2010) fluctuation test statistic
Figure B.13: Overall degree of shrinkage toward the prior mean for US inflation.
Figure B.14: RT and RW predictive densities
Figure B.15: Evolution of a quantile-based skewness measure for the RT predictive densities

REFERENCES

Aastveit, K. A., J. L. Cross, and H. K. van Dijk, "Quantifying Time-Varying Forecast Uncertainty and Risk for the Real Price of Oil," *Journal of Business & Economic Statistics* 41 (2023), 523–37. https://doi.org/10.1080/07350015.2022.2039159.

———, K. R. Gerdrup, A. S. Jore, and L. A. Thorsrud, "Nowcasting GDP in Real Time: A Density Combination Approach," *Journal of Business & Economic Statistics* 32 (2014), 48–68. https://doi.org/10.1080/07350015.2013.844155.

———, J. Mitchell, F. Ravazzolo, and H. K. van Dijk, "The Evolution of Forecast Density Combinations in Economics," in *Oxford Research Encyclopedia of Economics and Finance* (Oxford University Press, 2019). https://doi.org/10.1093/acrefore/9780190625979.013.381

———, F. Ravazzolo, and H. K. van Dijk, "Combined Density Nowcasting in an Uncertain Economic Environment," *Journal of Business & Economic Statistics* 36 (2018), 131–45. https://doi.org/10.1080/07350015.2015.1137760

Adrian, T., N. Boyarchenko, and D. Giannone, "Vulnerable Growth," *American Economic Review* 109 (2019), 1263–89. https://doi.org/10.1257/aer.20161923.

Baker, S., N. Bloom, and S. Davis, "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics* 131 (2016), 1593–636. https://doi.org/10.1093/qje/qjw024.

Bassetti, F., R. Casarin, and M. Del Negro, "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 R. Bachmann, G. Topa, and W. van der Klaauw, eds., *Handbook of Economic Expectations* (Elsevier, 2023), 443–476. https://doi.org/10.1016/B978-0-12-822927-9.00023-9.

———, ———, and F. Ravazzolo, "Bayesian Nonparametric Calibration and Combination of Predictive Distributions," *Journal of the American Statistical Association* 113 (2018), 675–85. https://doi.org/10.1080/01621459.2016.1273117.

Bates, J., and C. Granger, "The Combination of Forecasts," *Journal of the Operational Research Society* 20 (1969), 451–68. https://doi.org/10.1057/jors.1969.103.

Billio, M., R. Casarin, ———, and ———, "Time-Varying Combinations of Predictive Densities Using Nonlinear Filtering," *Journal of Econometrics* 177 (2013), 213–32. https://doi.org/10.1016/j.jeconom.2013.04.009.

Čapek, J., J. C. Cuaresma, N. Hauzenberger, and V. Reichel, "Macroeconomic Forecasting in the Euro Area Using Predictive Combinations of DSGE Models," *International Journal of Forecasting* 39 (2023), 1820–38. https://doi.org/10.1016/j.ijforecast.2022.09.002.

Casarin, R., S. Grassi, F. Ravazzolo, and ———, "A Flexible Predictive Density Combination for Large Financial Data Sets in Regular and Crisis Periods," *Journal of Econometrics* 237 (2023), 105370. https://doi.org/10.1016/j.jeconom.2022.11.004.

Chernis, T., "Combining Large Numbers of Density Predictions with Bayesian Predictive Synthesis," *Studies in Nonlinear Dynamics & Econometrics* 28(2) (2023), 293–317. https://doi.org/10.1515/snde-2022-0108.

———, G. Koop, E. Tallman, and M. West, "Decision Synthesis in Monetary Policy," Staff Working Papers 24-30, Bank of Canada, August 2024. https://doi.org/10.34989/swp-2024-30.

———, and T. Webley, "Nowcasting Canadian GDP with Density Combinations," Staff Discussion Papers 2022-12, Bank of Canada, May 2022. https://doi.org/10.34989/sdp-2022-12.

Chipman, H. A., E. I. George, and R. E. McCulloch, "Bayesian CART Model Search," *Journal of the American Statistical Association* 93 (1998), 935–48. https://doi.org/10.2307/2669832.

———, ———, and ———, "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics* 4 (2010), 266–98. https://doi.org/10.1214/09-AOAS285.

Clark, T. E., "Real-Time Density Forecasts from Bayesian Vector Autoregressions With Stochastic Volatility," *Journal of Business & Economic Statistics* 29 (2011), 327–41. https://doi.org/10.1198/jbes.2010.09248.

———, F. Huber, G. Koop, and M. Marcellino, "Forecasting US Inflation Using Bayesian Nonparametric Models," *The Annals of Applied Statistics* 18 (2024), 1421–44. https://doi.org/10.1214/23-AOAS1841.

———, ———, ———, ———, and M. Pfarrhofer, "Tail Forecasting with Multivariate Bayesian Additive Regression Trees," *International Economic Review* 64 (2023), 979–1022. https://doi.org/10.1111/iere.12619.

Conflitti, C., C. De Mol, and D. Giannone, "Optimal Combination of Survey Forecasts," *International Journal of Forecasting* 31 (2015), 1096–103. https://doi.org/10.1016/j.ijforecast.2015.03.009.

Creal, D., and J. Kim, "Empirical Asset Pricing with Bayesian Regression Trees," Technical Report, Working Paper University of Notre Dame, Department of Economics, 2021.

Coulombe, P. G., "The Macroeconomy as a Random Forest," *Journal of Applied Econometrics* 39(3) (2024), 401–21. https://doi.org/10.1002/jae.3030.

Del Negro, M., R. B. Hasegawa, and F. Schorfheide, "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance," *Journal of Econometrics* 192 (2016), 391–405. https://doi.org/10.1016/j.jeconom.2016.02.006.

Deshpande, S. K., R. Bai, C. Balocchi, J. E. Starling, and J. Weiss, "VCBART: Bayesian Trees for Varying Coefficients," *Bayesian Analysis* (2024), 1–28. https://doi.org/10.1214/24-BA1470.

Diebold, F. X., and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics* 13 (1995), 253–63. https://doi.org/10.1198/073500102753410444.

———, M. Shin, and B. Zhang, "On the Aggregation of Probability Assessments: Regularized Mixtures of Predictive Densities for Eurozone Inflation and Real Interest Rates," *Journal of Econometrics* 237 (2023), 105321. https://doi.org/10.1016/j.jeconom.2022.06.008.

ECB, "Results of the Third Special Questionnaire for Participants in the ECB Survey of Professional Forecasters," Technical Report, European Central Bank, February 2019. https://www.ecb.europa.eu/stats/ecbsurveys/surveyofprofessionalforecasters/html/ecb.spf201902specialsurvey7275f9e7e6.en.html.

Farrell, M. H., T. Liang, and S. Misra, “Deep Learning for Individual Heterogeneity: An Automatic Inference Framework,” *arXiv preprint arXiv:2010.14694* (2020). https://doi.org/10.48550/arXiv.2010.14694.

———, ———, and ———, “Deep Neural Networks for Estimation and Inference,” *Econometrica* 89 (2021), 181–213. https://doi.org/10.3982/ECTA16901.

Frühwirth-Schnatter, S., and H. Wagner, “Stochastic Model Specification Search for Gaussian and Partial Non-Gaussian State Space Models,” *Journal of Econometrics* 154 (2010), 85–100. https://doi.org/10.1016/j.jeconom.2009.07.003.

Garcia, J. A., “An Introduction to the ECB’s Survey of Professional Forecasters,” Occasional Paper Series 8, European Central Bank, September 2003. https://econpapers.repec.org/paper/ecbecbops/20038.htm.

Genest, C., and J. V. Zidek, “Combining Probability Distributions: A Critique and an Annotated Bibliography,” *Statistical Science* 1 (1986), 114–35. https://doi.org/10.1214/ss/1177013825

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

———, and G. Amisano, “Optimal Prediction Pools,” *Journal of Econometrics* 164 (2011), 130–41. https://doi.org/10.1016/j.jeconom.2011.02.017.

Giacomini, R., and B. Rossi, “Forecast Comparisons in Unstable Environments,” *Journal of Applied Econometrics* 25 (2010), 595–620. https://doi.org/10.1002/jae.1177.

Gneiting, T., and R. Ranjan, “Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules,” *Journal of Business & Economic Statistics* 29 (2011), 411–22. https://doi.org/10.1198/jbes.2010.08110.

Granger, C. W. J., and R. Ramanathan, “Improved Methods of Combining Forecasts,” *Journal of Forecasting* 3 (1984), 197–204. https://doi.org/10.1002/for.3980030207.

Hall, S. G., and J. Mitchell, “Combining Density Forecasts,” *International Journal of Forecasting* 23 (January 2007), 1–13. https://doi.org/10.1016/j.ijforecast.2006.08.001.

Hauzenberger, N., F. Huber, G. Koop, and J. Mitchell, “Bayesian Modeling of Time-Varying Parameters Using Regression Trees,” Technical Report, Federal Reserve Bank of Cleveland WP No. 23-05, 2023. https://doi.org/10.26509/frbc-wp-202305.

———, ———, ———, and L. Onorante, “Fast and Flexible Bayesian Inference in Time-Varying Parameter Regression Models,” *Journal of Business & Economic Statistics* 40 (2022), 1904–18. https://doi.org/10.1080/07350015.2021.1990772.

———, ———, M. Marcellino, and N. Petz, “Gaussian Process Vector Autoregressions and Macroeconomic Uncertainty,” *Journal of Business & Economic Statistics* 43(1) (2025), 27–43. https://doi.org/10.1080/07350015.2024.2322089.

Ho, P., T. A. Lubik, and C. Matthes, “Averaging Impulse Responses Using Prediction Pools,” *Journal of Monetary Economics* 146 (2024), 103571. https://doi.org/10.1016/j.jmoneco.2024.103571.

Huber, F., G. Koop, L. Onorante, M. Pfarrhofer, and J. Schreiner, “Nowcasting in a Pandemic Using Non-Parametric Mixed Frequency VARs,” *Journal of Econometrics* 232 (2023), 52–69. https://doi.org/10.1016/j.jeconom.2020.11.006.

———, and L. Rossini, “Inference in Bayesian Additive Vector Autoregressive Tree Models,” *The Annals of Applied Statistics* 16 (2022), 104–23. https://doi.org/10.1214/21-AOAS1488.

Jin, X., J. M. Maheu, and Q. Yang, “Infinite Markov Pooling of Predictive Distributions,” *Journal of Econometrics* 228 (2022), 302–21. https://doi.org/10.1016/j.jeconom.2021.10.010.

Kapetanios, G., J. Mitchell, S. Price, and N. Fawcett, “Generalised Density Forecast Combinations,” *Journal of Econometrics* 188 (2015), 150–65. https://doi.org/10.1016/j.jeconom.2015.02.047.

Knotek, E. S., and S. Zaman, “Real-Time Density Nowcasts of US Inflation: A Model Combination Approach,” *International Journal of Forecasting* 39 (October 2023), 1736–60. https://doi.org/10.1016/j.ijforecast.2022.04.007.

Koop, G., and D. Korobilis, “Forecasting Inflation Using Dynamic Model Averaging,” *International Economic Review* 53 (2012), 867–86. https://doi.org/10.1111/j.1468-2354.2012.00704.x.

Li, L., Y. Kang, and F. Li, “Bayesian Forecast Combination Using Time-Varying Features,” *International Journal of Forecasting* 39 (2023), 1287–302. https://doi.org/10.1016/j.ijforecast.2022.06.002.

Lopez-Salido, D., and F. Loria, “Inflation at Risk,” *Journal of Monetary Economics* 145 (2024), 103570. https://doi.org/10.1016/j.jmoneco.2024.103570.

McAlinn, K., K. A. Aastveit, J. Nakajima, and M. West, “Multivariate Bayesian Predictive Synthesis in Macroeconomic Forecasting,” *Journal of the American Statistical Association* 115 (2020), 1092–110. https://doi.org/10.1080/01621459.2019.1660171.

———, and M. West, “Dynamic Bayesian Predictive Synthesis in Time Series Forecasting,” *Journal of Econometrics* 210 (2019), 155–69. https://doi.org/10.1016/j.jeconom.2018.11.010.

McCracken, M. W., and S. Ng, “FRED-QD: A Quarterly Database for Macroeconomic Research,” *Review* 103 (January 2021), 1–44. https://doi.org/10.20955/r.103.1-44.

Mitchell, J., and S. G. Hall, "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 (2005), 995–1033. https://doi.org/10.1111/j.1468-0084.2005.00149.x.

Oelrich, O., M. Villani, and S. Ankargren, "Local Prediction Pools," *Journal of Forecasting* 43(1) (2023), 103–17. https://doi.org/10.1002/for.3030.

Rossi, B., and T. Sekhposyan, "Evaluating Predictive Densities of US Output Growth and Inflation in a Large Macroeconomic Data Set," *International Journal of Forecasting* 30 (2014), 662–82. https://doi.org/10.1016/j.ijforecast.2013.03.005

———, and ———, "Alternative Tests for Correct Specification of Conditional Predictive Densities," *Journal of Econometrics* 208 (2019), 638–57. https://doi.org/10.1016/j.jeconom.2018.07.008.

Stock, J. H., and M. W. Watson, "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature* 41 (2003), 788–829. https://doi.org/10.1257/002205103322436197.

———, and ———, "Why Has U.S. Inflation Become Harder to Forecast?" *Journal of Money, Credit and Banking* 39 (2007), 3–33. https://doi.org/10.1111/j.1538-4616.2007.00014.x.

Tallman, E., and M. West, "Bayesian Predictive Decision Synthesis," *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86 (10 2023), 340–63. https://doi.org/10.1093/jrsssb/qkad109.

Timmermann, A., "Forecast Combinations," in G. Elliott, C. W. Granger and A. Timmermann, eds., *Handbook of Economic Forecasting*, Volume 1 (Elsevier, 2006), 135–96. https://doi.org/10.1016/S1574-0706(05)01004-9.

Waggoner, D. F., and T. Zha, "Confronting Model Misspecification in Macroeconomics," *Journal of Econometrics* 171 (2012), 167–84. https://doi.org/10.1016/j.jeconom.2012.06.013.

Wallis, K. F., "Combining Density and Interval Forecasts: A Modest Proposal," *Oxford Bulletin of Economics and Statistics* 67 (2005), 983–94. https://doi.org/10.1111/j.1468-0084.2005.00148.x.

West, M., "Modelling Agent Forecast Distributions," *Journal of the Royal Statistical Society: Series B* 54 (January 1992), 553–67. https://doi.org/10.1111/j.2517-6161.1992.tb01896.x.

———, and J. Crosse, "Modelling Probabilistic Agent Opinion," *Journal of the Royal Statistical Society: Series B* 54 (September 1992), 285–99. https://doi.org/10.1111/j.2517-6161.1992.tb01882.x

Woodford, M., "The Case for Forecast Targeting as a Monetary Policy Strategy," *Journal of Economic Perspectives* 21 (December 2007), 3–24. https://doi.org/10.1257/jep.21.4.3