DISCUSSION PAPER SERIES

DP17461

Tail Forecasting with Multivariate Bayesian Additive Regression Trees

Todd Clark, Florian Huber, Gary Koop, Massimiliano Marcellino and Michael Pfarrhofer

MONETARY ECONOMICS AND FLUCTUATIONS

ISSN 0265-8003

Tail Forecasting with Multivariate Bayesian Additive Regression Trees

Todd Clark, Florian Huber, Gary Koop, Massimiliano Marcellino and Michael Pfarrhofer

Discussion Paper DP17461
Published 12 July 2022
Submitted 08 July 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Monetary Economics and Fluctuations

Tail Forecasting with Multivariate Bayesian Additive Regression Trees

Abstract

We develop multivariate time series models using Bayesian additive regression trees that posit nonlinearities among macroeconomic variables, their lags, and possibly their lagged errors. The error variances can be stable, feature stochastic volatility, or follow a nonparametric specification. We evaluate density and tail forecast performance for a set of US macroeconomic and financial indicators. Our results suggest that the proposed models improve forecast accuracy both overall and in the tails. Another finding is that when allowing for nonlinearities in the conditional mean, heteroskedasticity becomes less important. A scenario analysis reveals nonlinear relations between predictive distributions and financial conditions.

JEL Classification: C11, C32, C53

Keywords: Nonparametric VAR, regression trees, Macroeconomic forecasting, Scenario analysis

Todd Clark - todd.clark@researchfed.org
Federal Reserve Bank of Cleveland

Florian Huber - florian.huber@plus.ac.at
University of Salzburg

Gary Koop - gary.koop@strath.ac.uk
Strathclyde University

Massimiliano Marcellino - massimiliano.marcellino@unibocconi.it
Bocconi University and CEPR

Michael Pfarrhofer - michael.pfarrhofer@plus.ac.at
University of Salzburg

Tail Forecasting with Multivariate Bayesian Additive Regression Trees

Todd E. Clark
Federal Reserve Bank of Cleveland, United States

Florian Huber[1]
University of Salzburg, Austria

Gary Koop
University of Strathclyde, United Kingdom

Massimiliano Marcellino
Bocconi University, IGIER, Baffi, Bidsa, CEPR, Italy

Michael Pfarrhofer
University of Salzburg, Austria

We develop multivariate time series models using Bayesian additive regression trees that posit nonlinearities among macroeconomic variables, their lags, and possibly their lagged errors. The error variances can be stable, feature stochastic volatility, or follow a nonparametric specification. We evaluate density and tail forecast performance for a set of US macroeconomic and financial indicators. Our results suggest that the proposed models improve forecast accuracy both overall and in the tails. Another finding is that when allowing for nonlinearities in the conditional mean, heteroskedasticity becomes less important. A scenario analysis reveals nonlinear relations between predictive distributions and financial conditions.

JEL: C11, C32, C53

KEYWORDS: Nonparametric VAR, regression trees, macroeconomic forecasting, scenario analysis

[1]The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Federal Reserve System. Marcellino thanks MIUR – PRIN Bando 2017 – prot. 2017TA7TYC for financial support; Huber and Pfarrhofer gratefully acknowledge financial support from the Austrian Science Fund (FWF, grant no. ZK 35). We would like to thank the editor Jesus Fernandez-Villaverde, three anonymous referees, Luc Bauwens, Josh Chan, Herman van Dijk, Gernot Doppelhofer, Christian Hafner, Ivan Jeliazkov, Mike McCracken, James Mitchell, Luca Onorante, Mike West, and seminar participants at the Centre for Operations Research and Econometrics, UC Irvine, University of Sydney, Norwegian School of Economics, the 2021 IAAE conference, 11th ECB Conference on Forecasting Techniques, and the European Seminar on Bayesian Econometrics 2021 for helpful comments. A previous version of this paper was circulated as Federal Reserve Bank of Cleveland Working Paper 21-08. Please address correspondence to: Florian Huber. Department of Economics, University of Salzburg. *Address*: Mönchsberg 2a, 5020 Salzburg, Austria. *Email*: florian.huber@plus.ac.at.

1 Introduction

Two recent events, the global financial crisis and the COVID-19 pandemic, have increased interest in tail risks in macroeconomic outcomes. A fast-growing literature has focused on the risks of significant declines in GDP, with quantile regression being the main method used to estimate tail risks (see, e.g., Adrian, Boyarchenko, and Giannone (2019); Adrian, et al. (2022); Cook and Doh (2019); Delle Monache, De Polis, and Petrella (2020); De Nicolò and Lucchetta (2017); Ferrara, Mogliani, and Sahuc (2022); Giglio, Kelly, and Pruitt (2016); González-Rivera, Maldonado, and Ruiz (2019); Mitchell, Poon, and Mazzi (2022); Plagborg-Møller, et al. (2020); and Reichlin, Ricco, and Hasenzagl (2020)).[2] Some studies focused on tail risks to unemployment (e.g., Galbraith and van Norden (2019) and Kiley (2022)) or inflation (e.g., Ghysels, Iania, and Striaukas (2018)) or deal with forecasting the complete distribution of several macroeconomic aggregates (see Manzan (2015), Korobilis (2017), and Manzan and Zerom (2013, 2015)).

The present paper departs from this literature by using Bayesian parametric and nonparametric time series models instead of quantile regression. Our focus is motivated by Carriero, Clark, and Marcellino (forthcoming, CCM), who evaluate the ability of alternative econometric methods to produce accurate nowcasts of tail risks to GDP growth, possibly in the presence of a large information set. They find that Bayesian quantile regression performs much better than classical quantile regression and that Bayesian linear regression performs similarly or sometimes better for tail forecasting, once endowed with stochastic volatility (SV).[3]

A parallel, and also fast-growing, literature evaluates the use of machine learning techniques for macroeconomic forecasting, with random forests (see Breiman (2001) and, e.g., Masini, Medeiros, and Mendes (2021), for a survey) performing particularly well, also during crisis times, in a variety of studies and for key variables such as GDP growth and inflation; see, e.g., Goulet Coulombe (2020), Goulet Coulombe, et al. (2020); Goulet Coulombe, Marcellino, and

[2]For output growth, forecasting tail risks has some precedent in the literature on forecasting recessions or just periods of negative growth (see, e.g., Aastveit, Ravazzolo, and van Dijk (2018)).

[3]The intuition for this finding, explained more formally in Carriero, Clark, and Marcellino (2020), is that the explanatory variables drive changes in the conditional mean of growth, which decreases during crisis times, while SV permits an increase in the conditional variance. Thus, the left tail of the conditional distribution of growth can decrease more than the right tail during crisis times, generating the kind of asymmetries emphasized in the quantile regression-based literature. Caldara, Scotti, and Zhong (2021) make a similar point using a model with leverage, where the estimated SV enters the conditional mean with a negative coefficient.

Stevanovic (2021), and Medeiros, et al. (2021). While these papers adopt classical methods, Bayesian techniques are also available. In particular, Bayesian additive regression trees (BART; see Chipman, George, and McCulloch (2010)) provide a flexible approach popular in many fields of statistics. Huber and Rossini (2022, HR) develop Bayesian methods that build BART into a vector autoregression (VAR), leading to the Bayesian additive vector autoregressive tree model, and demonstrate that it forecasts well. Huber, et al. (2020, HKOPS) develop Bayesian methods for the mixed-frequency version of this model, showing that it also forecasts well, particularly during the COVID-19 pandemic.

In this paper, we combine the tail forecasting focus of CCM with the BART methodology of HR and HKOPS. The first contribution of this paper is methodological and lies in the development of a set of novel and easy to use nonparametric econometric models that can be applied in a variety of contexts. Specifically, in addition to the original model of HR (we use the acronym BART for this model), we introduce three novel alternative BART-based nonparametric VARs that, we argue, have properties that make them potentially useful for empirical macroeconomic modeling and forecasting, particularly in unstable times. These competing specifications arise from a general nonparametric multivariate regression model by choosing suitable covariates that can be observed or (partially) latent. Each of these nested models has different implications for how a given model treats nonlinearities in different regions of the parameter space.

The flexible modeling of the conditional mean in BART-based specifications could make the error variance more stable than in linear models, and we do consider homoskedastic versions of our nonparametric models. But this is not necessarily the case. Hence, we also focus on versions of our models complemented either with SV or with a novel nonparametric specification for the time variation in the conditional variance, related to that in Pratola, et al. (2020) and labeled heteroBART (hBART).

Our second contribution is the development of general Markov chain Monte Carlo (MCMC) estimation algorithms that are applicable to large-dimensional models. These methods are designed for homoskedastic and heteroskedastic models and build on a parsimonious factor structure in the shocks to permit fast estimation of large systems. To sample from the posterior of the hBART volatility model, we propose a novel updating step based on using the auxiliary sampler for SV models developed in Omori, et al. (2007).

Our final contribution is empirical. Using real-time data for a set of US macroeconomic and financial indicators, we first assess the performance of the various BART models for den-

sity and tail forecasting using several commonly used metrics of tail and density forecasting accuracy. The different nonparametric models are benchmarked to several popular models commonly used in the literature such as Bayesian VARs (BVARs) with SV (Clark (2011), Clark and Ravazzolo (2015), and Koop (2013)), a BVAR with time-varying parameters and SV (see, e.g., Huber, Koop, and Onorante (2021)), and a Bayesian quantile regression (BQR, see Kozumi and Kobayashi, 2011). After showing that BART-based models improve upon the competing models (especially so at longer forecast horizons), we drill deeper into the properties of the predictive distributions of the best-performing BART specification. In addition, we also illustrate how the model can be used to carry out conditional forecasts to analyze the relationship between the macroeconomy and financial conditions.

The empirical results can be summarized as follows. BART-based models improve upon the competing models in terms of joint density forecasting performance. These performance gains are especially pronounced for higher-order forecasts. Accuracy improvements are mostly driven by superior point forecasts and higher-order features of the predictive distributions, with the former being more important than the latter. When the focus is on tail forecasting, our models display a similarly strong performance, suggesting that taking into account nonlinearities in the conditional mean is relevant for producing precise tail forecasts. Once we use a flexible conditional mean model, controlling for heteroskedasticity becomes less of a concern. Finally, conditioning on different values of the national financial conditions index reveals highly nonlinear interactions between one-step-ahead predictive distributions and financial conditions.

The paper proceeds as follows. Section 2 describes the various multivariate Bayesian additive regression tree models and Section 3 discusses Bayesian inference. Section 4 considers the data, forecast design, and evaluation metrics used in the empirical application and discusses empirical findings. Section 5 summarizes and concludes. Additional empirical results are included in an appendix.

2 Nonparametric modeling of VARs using BART

This section explains the BART formulations considered in this paper. In a multivariate time series model such as a VAR, specification choices are made for conditional means and conditional variances. For instance, in the classic BVAR-SV model the conditional means are linear and log conditional variances follow random walks. In this paper, we compare this model to various models that are partially or completely nonparametric. In various combinations, the models

include parametric and nonparametric representations of the conditional mean of a VAR as well as of the conditional variance.

2.1 Nonparametric VARs

Let $\{\boldsymbol{y}_t\}_{t=1}^T$ denote an M-dimensional vector of macroeconomic and financial time series with typical i^{th} element y_{it}. We assume that $\boldsymbol{y}_t$ depends on its p lags, which we store in a $K(=Mp)$-dimensional vector $\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \dots, \boldsymbol{y}'_{t-p})'$. The relationship between $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$ is assumed to be unknown and potentially highly nonlinear. This is captured through the following general multivariate model:

$$\boldsymbol{y}_t = F(\boldsymbol{x}_t) + \boldsymbol{\eta}_t, \tag{1}$$

$$\boldsymbol{\eta}_t = G(\boldsymbol{z}_t) + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{\Sigma}_t). \tag{2}$$

Here, we let $F : \mathbb{R}^K \to \mathbb{R}^M$ and $G : \mathbb{R}^N \to \mathbb{R}^M$ denote unknown functions with $F(\boldsymbol{x}_t) = (f_1(\boldsymbol{x}_t), \dots, f_M(\boldsymbol{x}_t))'$ and $G(\boldsymbol{z}_t) = (g_1(\boldsymbol{z}_t), \dots, g_M(\boldsymbol{z}_t))'$, while f_j and g_j are equation-specific scalar-valued functions. $\boldsymbol{z}_t$ is a vector of additional explanatory variables with dimension $N \times 1$ that is defined below in the context of our different models.

Although we also consider conditionally homoskedastic implementations, we generally treat the shocks in $\boldsymbol{\varepsilon}_t$ as following a multivariate Gaussian distribution with a time-varying variance-covariance matrix $\boldsymbol{\Sigma}_t$. To ensure parsimony and allow for fast estimation, we will assume that $\boldsymbol{\Sigma}_t$ features factor stochastic volatility (FSV; see Aguilar and West (2000)):

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}\boldsymbol{\Omega}_t\boldsymbol{\Lambda}' + \boldsymbol{H}_t \quad \Longleftrightarrow \quad \boldsymbol{\varepsilon}_t = \boldsymbol{\Lambda}\boldsymbol{\delta}_t + \boldsymbol{e}_t, \quad \boldsymbol{\delta}_t \sim \mathcal{N}(\boldsymbol{0}_Q, \boldsymbol{\Omega}_t), \quad \boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{H}_t),$$

with $\boldsymbol{\Lambda}$ denoting an $M \times Q$ matrix of factor loadings (with $Q \ll M$) and diagonal variance-covariance matrices $\boldsymbol{\Omega}_t = \text{diag}(e^{u_1(\boldsymbol{w}_t)}, \dots, e^{u_Q(\boldsymbol{w}_t)})$ and $\boldsymbol{H}_t = \text{diag}(e^{v_1(\boldsymbol{w}_t)}, \dots, e^{v_M(\boldsymbol{w}_t)})$ with $u_i, v_j : \mathbb{R}^R \to \mathbb{R}$ being unknown functions that describe how the error variances are related to a set of R covariates in $\boldsymbol{w}_t$. Conditionally on the Q factors $\boldsymbol{\delta}_t$, the shocks in $\boldsymbol{e}_t$ are uncorrelated and the model can be estimated on an equation-by-equation basis. This factor specification is, without further restrictions, not identified. Since our focus is on tail forecasting and we do not aim to interpret $\boldsymbol{\Lambda}$ and $\boldsymbol{\delta}_t$ separately but are exclusively interested in $\boldsymbol{\Sigma}_t$, this causes no additional issues. Suitable identification schemes can be straightforwardly introduced. For the

law of motion of u_i, v_j we will use both a standard SV model and a more flexible specification closely related to the heteroskedastic BART model proposed in Pratola, et al. (2020). More details are provided in Section 2.4.

To make this model operational we need to learn the unknown functions F and G. We will discuss how we do this in the next subsection.

2.2 Function learning using BART

BART approximates the unknown functions F and G using a sum of regression trees. In what follows, our focus will be on estimating the function associated with the i^{th} equation.

Let $\boldsymbol{y}_{i\bullet}$ denote the i^{th} column of $\boldsymbol{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_T')'$, $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_T')'$, $\boldsymbol{Z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_T')'$ (with $\boldsymbol{x}_{\bullet j}$ denoting the j^{th} column of $\boldsymbol{X}$), and $\boldsymbol{\varepsilon}_{\bullet i} = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$. The i^{th} equation of the nonlinear VAR outlined in Eqs. (1) and (2) is:

$$\boldsymbol{y}_{i\bullet} = f_i(\boldsymbol{X}) + g_i(\boldsymbol{Z}) + \boldsymbol{\varepsilon}_{\bullet i}. \tag{3}$$

BART replaces f_i and g_i with a sum-of-trees approximation:

$$f_i(\boldsymbol{X}) \approx \sum_{s=1}^{S} l_{is}(\boldsymbol{X}|\mathcal{T}_{is}^f, \boldsymbol{\mu}_{is}^f), \quad g_i(\boldsymbol{Z}) \approx \sum_{s=1}^{S} l_{is}(\boldsymbol{Z}|\mathcal{T}_{is}^g, \boldsymbol{\mu}_{is}^g).$$

Here, we let l_{is} denote a (single) regression tree function. A tree is a step function that can be defined as follows (see Ročková and Saha, 2019):

$$l_{is}(\boldsymbol{X}|\mathcal{T}_{is}^j, \boldsymbol{\mu}_{is}^f) = \sum_{n=1}^{b_{is}^j} \mu_{is,n}^j \mathbb{I}(\boldsymbol{X} \in \mathcal{S}_{is,n}^j), \tag{4}$$

with $\mathbb{I}(\bullet)$ denoting an indicator function that equals 1 if its argument is true, $\mathcal{S}_{is,n}^j$ is a set associated with a terminal node (for $j \in \{f, g\}$), and $\boldsymbol{\mu}_{is}^j = \left(\mu_{is,1}^j, \ldots, \mu_{is,b_{is}^j}^j\right)'$ are terminal node parameters of dimension b_{is}^j.

The tree function is fully determined by the terminal node parameters $\boldsymbol{\mu}_{is}^j$ and the tree structure $\mathcal{T}_{is}^j = \{\mathcal{S}_{is,n}^j\}_{n=1}^{b_{is}^j}$. The partitions are binary, implying that each node can split into two child nodes (and thus act as an interior node) or end up being a terminal node. Each set $\mathcal{S}_{is,n}^j$ is constructed by partitioning the data according to splitting rules of the form $\{x_{\bullet j} \leq c\}$ or $\{x_{\bullet j} > c\}$ for $j = 1, \ldots, K$ (in the case of approximating f). The threshold c takes a value from the range of the observed values $\{x_{jt}\}_{t=1}^T$. The splitting rules in the case of $\mathcal{T}_{is}^g$ are defined

analogously.

The number of trees S should be large to achieve a certain degree of representation flexibility, while Bayesian shrinkage priors are used to avoid overfitting. Chipman, George, and McCulloch (2010) show that, using 42 different data sets, the choice of S is only weakly influential as long as S is set larger than 50. They recommend a value of $S = 200$. Huber, et al. (2020), using macroeconomic time series data, reinforce this finding. In particular, they show that predictive performance improves with S as long as S is below 50. Once S exceeds 50 the forecast accuracy (measured through log predictive likelihoods) does not change substantially.

Before we illustrate BART using a simple example, it is worth discussing why BART is nonparametric. In general, the key idea of nonparametric inference is to estimate an unknown quantity (such as the functions F and G) making as few assumptions as possible. Wasserman (2006, p. 12) loosely defines nonparametric inference as "a set of modern statistical methods that aim to keep the number of underlying assumptions as weak as possible." Since BART builds on few assumptions to infer the unknown functions, it fulfills this definition.

A second definition of a nonparametric model that is frequently used states that, as opposed to parametric inference, the number of parameters is not fixed a priori and adapts to the complexity of the data set. To see this, notice that the topology of the tree structure is determined by a sequence of recursive decision rules. These decision rules imply that the dimension of the vector of terminal node parameters is not fixed a priori and depends on the size of the input data set. This immediately implies that BART automatically adapts to the complexity of the data and that the dimension of the parameter space grows with the number of observations T.

The third definition is that the parameter space is infinite dimensional. We will use a Bayesian prior that governs how trees are generated. This process, described below, can be viewed as a prior over the infinite dimensional space of possible functions F and G. However, similar to infinite mixtures and Gaussian processes, once we condition on the data the model becomes finite dimensional, permitting estimation.

To illustrate what BART does, we focus on a special case of Eq. (3) assuming a single tree ($S = 1$). To simplify the notation, we drop the tree and equation-specific subscripts i and s as well as the superscript j. The corresponding regression tree model is then given by:

$$\boldsymbol{y} = l(\boldsymbol{X}|\mathcal{T}, \boldsymbol{\mu}) + \boldsymbol{\varepsilon}.$$

The conditional mean of this model is similar to Eq. (4):

$$E(\boldsymbol{y}|\boldsymbol{x}) = l(\boldsymbol{X}|\mathcal{T}, \boldsymbol{\mu}) = \sum_{n=1}^{b} \mu_n \mathbb{I}(\boldsymbol{X} \in \mathcal{S}_n).$$

This equation suggests that the conditional mean under a single tree is a piece-wise constant function that assigns μ_n if a specific configuration of $\boldsymbol{X}$ is in the set $\mathcal{S}_n$. Notice that this is a simple analysis of variance (ANOVA) model that can be stated in terms of a multivariate regression model conditional on the indicators.

In the case where the tree is simple (i.e., if b is small), the corresponding conditional mean will feature a relatively low number of breaks. Hence, such a model explains only a small fraction of the variation in $\boldsymbol{y}$. In the machine learning literature, this is often referred to as a weak learner. Instead of fitting more complex trees, BART builds on the notion that summing over many simple trees (which are pruned using Bayesian shrinkage) improves upon using a single complex tree.[4] The resulting conditional mean, when the trees are viewed together, allows for capturing rich dynamics in $\boldsymbol{y}$, implying strong explanatory power. In this case regularization helps to avoid issues related to overfitting.

Before we proceed, one word on statistical identification of BART is necessary. Since we sum over many trees, each individual tree is, strictly speaking, not identified. That is, different trees can lead to the same function. This, however, poses no issues, since our interest does not center on a specific tree but only on the sum over all the trees. In fact, Chipman, George, and McCulloch (2010) notice that the abundance of unidentified parameters even improves MCMC mixing.

2.3 Nested model specifications

In terms of modeling the conditional mean, the specific choices made for F, G, and $\boldsymbol{z}_t$ allow for a wide range of flexible models. We discriminate between models that assume that either F or G (or both) is unknown and potentially nonlinear functions. The key notion is that $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ differ in the way they impact $\boldsymbol{y}_t$.

In this paper, we focus on four different model specifications that differ in the choice of F, G, and $\boldsymbol{z}_t$. The first model is a multivariate nonparametric VAR model. Assuming that

[4] This is closely related to ensemble methods and model averaging techniques that combine models to obtain more precise forecasts (see the discussion in Hernández, et al. (2018)).

$G(\boldsymbol{z}_t) = \mathbf{0}_M$ for all t, the model in Eq. (1) and Eq. (2) reduces to:

$$\boldsymbol{y}_t = F(\boldsymbol{x}_t) + \boldsymbol{\varepsilon}_t,$$

which posits a nonlinear relationship between $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$ and no effect of $\boldsymbol{z}_t$ on $\boldsymbol{y}_t$. Since F is approximated using BART we obtain the model proposed in Huber and Rossini (2022) and applied to the mixed-frequency case in Huber, et al. (2020). In the remainder of the paper, this model is labeled the BART model.

The second model we propose assumes that $\boldsymbol{z}_t = \boldsymbol{x}_t$ and $G(\boldsymbol{x}_t)$ is unknown and nonlinear, while $F(\boldsymbol{x}_t)$ is linear and depends on an $M \times K$ coefficient matrix $\boldsymbol{A}$. The corresponding model reads:

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t + G(\boldsymbol{x}_t) + \boldsymbol{\varepsilon}_t, \tag{5}$$

which is a multivariate additive regression model that assumes that there exists a linear VAR part, $\boldsymbol{A}\boldsymbol{x}_t$, and some unknown nonlinear part, $G(\boldsymbol{x}_t)$, which we approximate using BART. Intuitively speaking, this model assumes that the shocks $\boldsymbol{\eta}_t$ follow a nonlinear regression specification that serves to control for any nonlinear effects that persist after controlling for linear relations. In the remainder of the paper we label this the mixture BART (mixBART) model.

If we set $\boldsymbol{z}_t = (\boldsymbol{\eta}_{t-1}', \ldots, \boldsymbol{\eta}_{t-p}')'$ and $F(\boldsymbol{x}_t) = \boldsymbol{A}\boldsymbol{x}_t$, the resulting model implies that the reduced-form shocks $\boldsymbol{\eta}_t$ depend nonlinearly on their recent past. This specification allows for flexible adjustments of the conditional mean by exploiting information contained in past reduced-form shocks. During recessions such as that caused by the COVID-19 pandemic, this feature could help to quickly adjust forecasts in the presence of large historical forecast errors. Again, we use BART to approximate G, leading to the errorBART model.[5]

Finally, the last model we consider assumes that $\boldsymbol{\Sigma}_t = \boldsymbol{H}_t$, a diagonal matrix (i.e., $\boldsymbol{\Lambda} = \mathbf{0}_{M \times Q}$), implying that the shocks $\boldsymbol{\varepsilon}_t$ are independent. To capture possible nonlinear contemporaneous relations across equations we augment each equation with the shocks of the preceding equations. This gives rise to a model with a nonlinear covariance structure.

[5] Another feasible option, which would require approximation-based techniques along the lines used in Huber, et al. (2020), would be to specify $\boldsymbol{x}_t$ equal to the lags of $\boldsymbol{\eta}_t$ and $\boldsymbol{\varepsilon}_t$. This would be a nonparametric variant of a multivariate ARMA model (for a Bayesian treatment of ARMA models, see Chib and Greenberg (1994)).

The first equation of this model is:

$$y_{1t} = f_1(\boldsymbol{x}_t) + \varepsilon_{1t}.$$

The second equation depends nonlinearly on $\boldsymbol{x}_t$ and ε_{1t} as follows:

$$y_{2t} = f_2(\boldsymbol{x}_t) + g_2(\varepsilon_{1t}) + \varepsilon_{2t}.$$

In general, the i^{th} equation is given by:

$$y_{it} = f_i(\boldsymbol{x}_t) + g_i(\boldsymbol{r}_{it}) + \varepsilon_{it}, \tag{6}$$

with $\boldsymbol{r}_{it} = (\varepsilon_{1t}, \ldots, \varepsilon_{i-1,t})'$ being an $(i-1)$-dimensional vector of shocks. The model assumes that the contemporaneous relations across the shocks take a nonlinear form. This specification implicitly assumes a nonlinear variance-covariance structure. Modeling the covariances in a nonlinear way gives rise to (at least) two convenient features. First, the model is capable of quickly reacting to large shocks. Second, covariances are allowed to change over time, since different configurations of $\boldsymbol{r}_{it}$ can yield different fitted values.

Across all models considered, this specification provides the largest degree of flexibility, since it allows for a nonlinear mean function F as well as a nonlinear covariance function G with its argument differing across equations. In what follows, both F and G are again approximated using BART, leading to what we call the fullBART model.[6]

2.4 Adding heteroskedasticity to the model

Up to this point we have not discussed the specific functional forms of u_i and v_j or the choice of $\boldsymbol{w}_t$. In principle, we could set $\boldsymbol{w}_t = 1$ and simply assume u_i and v_j for all i, j to be linear functions. This would imply a homoskedastic model with a parsimonious specification for the error variance-covariance matrix that allows for fast estimation even in large systems. However,

[6]It is worth stressing that to estimate this model we use an equation-by-equation estimation MCMC algorithm based on augmenting each equation with the shocks of the previous equations. This approach closely resembles the algorithm put forth in Carriero, Clark, and Marcellino (2019) which is approximate in the sense that it does not sample from the exact full conditionals. Carriero, et al. (2022a) offer an exact algorithm that is not applicable in our model. Hence, we view this algorithm as an approximate one.

several recent papers have shown that allowing for conditional heteroskedasticity sharply improves density forecasts of macroeconomic aggregates (see, among others, Carriero, Clark, and Marcellino (2016); Clark (2011); Clark and Ravazzolo (2015)). In one set of models, we pair the BART formulations described above with conventional factor stochastic volatility of the innovations to the model. These models assume that the latent volatility process evolves according to a simple stochastic process that is persistent (in our implementation, an AR(1) model with a persistence parameter close to 1). During a pandemic, this high persistence in the volatility process could be detrimental for predictive accuracy, since the predictive variance only slowly adjusts to new information.[7]

These models allow for richer volatility dynamics but also assume a parametric and known law of motion. Accordingly, in another set of results for BART models, we propose an alternative volatility specification based on heteroskedastic BART (hBART, Pratola, et al. (2020)). The functions u_i and v_j are again approximated through BART:

$$u_i(\boldsymbol{w}) \approx \sum_{s=1}^{S} l_{is}(\boldsymbol{w}|\mathcal{T}_{is}^u, \boldsymbol{\mu}_{is}^u), \quad \text{for } i = 1, \dots, Q,$$
$$v_j(\boldsymbol{w}) \approx \sum_{s=1}^{S} l_{is}(\boldsymbol{w}|\mathcal{T}_{js}^v, \boldsymbol{\mu}_{js}^v), \quad \text{for } j = 1, \dots, M.$$

We call this model specification (factor) hBART since it assumes that the latent factors $\boldsymbol{\delta}_t$ and the measurement errors $\boldsymbol{e}_t$ are conditionally heteroskedastic with volatilities evolving according to a flexible BART specification. As compared to stochastic volatility models that feature separate shocks to determine the log-volatility process, our specification is closer to a (nonlinear) GARCH model (see, e.g., Bollerslev (1990); Sentana (1995); Engle (2002)), which implies a deterministic law of motion for the volatilities.

Selecting appropriate predictors $\boldsymbol{w}_t$ is crucial. In our empirical work, we consider $\boldsymbol{w}_t = (t, \boldsymbol{x}_t')'$. This choice has the advantage that our model allows for a (potentially) nonlinear trend and it assumes that the lagged values of $\boldsymbol{y}_t$ impact not only the conditional mean but also the error variances. Since the different decision rules might only depend on selected elements in $\boldsymbol{w}_t$, we do not risk overfitting if M or p is large. Moreover, and this turns out to be a key advantage, our choice of $\boldsymbol{w}_t$ allows for multi-step predictions of the error variances. More precisely, this is achieved by using Eqs. (1) and (2) to obtain a draw from the one-step-ahead

[7] As a solution, Carriero, et al. (2022b) discuss several alternative volatility models that allow for combining transitory and persistent changes in the volatility.

predictive distribution, labeled $\hat{\boldsymbol{y}}_{T+1}$, which is then used to compute $\boldsymbol{H}_{T+2}$ and $\boldsymbol{\Omega}_{T+2}$ based on $\boldsymbol{w}_{T+2} = (T+2, \hat{\boldsymbol{y}}'_{T+1}, \boldsymbol{y}'_T, \ldots, \boldsymbol{y}'_{T-p+1})'$. $\boldsymbol{H}_{T+2}$, in turn, allows us to generate a draw from the two-step-ahead predictive distribution, $\tilde{\boldsymbol{y}}_{T+2}$, which is based on $\boldsymbol{H}_{T+2}$ and $\boldsymbol{\Omega}_{T+2}$. In general, the h-step-ahead forecast distribution can be obtained analogously.

3 Bayesian inference

We estimate our model using Bayesian techniques. Although classical methods are available for some nonparametric models (see the studies cited in Section 1), the shrinkage that comes with Bayesian techniques is generally known to be helpful in macroeconomic forecasting (see, e.g., Carriero, Clark, and Marcellino (2015); Chan (2021); Giannone, Lenza, and Primiceri (2015); Huber and Feldkircher (2019); Stock and Watson (2012)). Our prior setup closely follows Huber, et al. (2020). Here, we focus on the prior associated with the tree structures $\mathcal{T}_{is}^j$ and the terminal node parameters $\boldsymbol{\mu}_{is}^j$. Chipman, George, and McCulloch (2010) build on Chipman, George, and McCulloch (1998) and propose a benchmark prior that induces shrinkage on the trees as well as on the terminal node parameters. We adopt this prior since it has been shown to work well for a wide variety of different data sets and for both in- and out-of-sample applications. Since the priors on the remaining coefficients are relatively standard, we provide additional information in Appendix B.

3.1 Priors on the trees and terminal node parameters

We do not specify a prior directly on the trees but instead design a tree-generating stochastic process that serves as a prior (see Chipman, George, and McCulloch (1998)). Ročková and Saha (2019) discuss how this stochastic process is linked to the Galton-Watson (GW) process, which models how a population of individuals reproduces dynamically. The trees can be viewed as the individuals who reproduce and die according to laws of chance.

This process features three aspects. The first is related to the probability that a node at depth $d = 1, \ldots,$ is nonterminal. Let $\alpha \in (0, 1)$ and $\beta \in \mathbb{R}^+$ be hyperparameters. The probability that a node at depth d gives rise to two child nodes is given by:

$$\frac{\alpha}{(1+d)^\beta}.$$

In our empirical work, we set $\alpha = 0.95$ and $\beta = 2$ for the trees $\mathcal{T}_{is}^j$ for all i, s, j. Chipman,

George, and McCulloch (2010) recommend these values for α and β as a standard choice that works well across a wide variety of different data sets. This prior implies that the probability that trees grow large decreases in d and thus favors smaller trees.

The second aspect of the prior is concerned with the selection of the variables that are used in a splitting rule. Here, we use a discrete uniform prior, which implies that we do not introduce prior information on which variables show up in a splitting rule. Finally, the third component is concerned with the specific value of the thresholds in the splitting rule. For these, we use a uniform prior over the range of the splitting variable as well.[8]

On the terminal node parameters, we use independent Gaussian priors that are specified as follows:

$$\mu^j_{is,k} \sim \mathcal{N}(0, \phi^j_{is,k}), \text{ for } k = 1, \ldots, b^j_{is}. \tag{7}$$

Following Chipman, George, and McCulloch (2010) we set the prior variance $\phi^j_{is,k}$ in a data-based way. The key idea is to specify the prior such that a certain amount of prior mass is placed on the range of the data but at the same time set the prior in a way that it introduces more shrinkage if S is large. A specification for $\phi^j_{is,k}$ that achieves this is:

$$\sqrt{\phi^j_{is,k}} = \frac{\max(\boldsymbol{z}^j_i) - \min(\boldsymbol{z}^j_i)}{2\gamma\sqrt{S}}, \tag{8}$$

where $\boldsymbol{z}^j_i$ is a T-dimensional vector that is equal to $\boldsymbol{z}^f_i = \boldsymbol{y}_{i\bullet} - g_i(\boldsymbol{Z}) - \boldsymbol{\delta}\boldsymbol{\lambda}'_i$ if $j = f$, $\boldsymbol{z}^g_i = \boldsymbol{y}_{i\bullet} - f_i(\boldsymbol{X}) - \boldsymbol{\delta}\boldsymbol{\lambda}'_i$ if $j = g$, $\boldsymbol{z}^v_i = \log((\boldsymbol{y}_{i\bullet} - f_i(\boldsymbol{X}) - g_i(\boldsymbol{Z}) - \boldsymbol{\delta}\boldsymbol{\lambda}'_i)^2)$, and $\boldsymbol{z}^u_k = \log(\boldsymbol{\delta}^2_k)$ with $\boldsymbol{\delta}_k$ denoting the k^{th} column of $\boldsymbol{\delta} = (\boldsymbol{\delta}'_1, \ldots, \boldsymbol{\delta}'_T)'$, and $\boldsymbol{\lambda}_i$ is the i^{th} row of $\boldsymbol{\Lambda}$. The parameter γ controls the tightness of the prior, with smaller values leading to a prior that puts more prior mass on the range of $\boldsymbol{z}^j_i$.

As noted by Huber, et al. (2020) this prior has the advantage of becoming increasingly loose (for fixed values of S and γ) if $\boldsymbol{z}^j_i$ includes outliers. This leads to a wider predictive distribution and thus a higher likelihood of observing outlying values. Chipman, George, and

[8] This tree-generating process prior can generate trees that are more complicated than necessary. For instance, in the case of a regression with a single binary explanatory variable, a trivial tree would be sufficient. The prior allows for trivial but also more complicated trees that could fit the data equally well. This property ensures that complicated trees are not ruled out a priori, allowing our posterior simulator to explore the space of trees efficiently. The parsimony built into the prior, however, will favor simple over complicated trees.

McCulloch (2010) propose $\gamma = 2$ in combination with transforming the data such that $\boldsymbol{z}_i^j$ ranges from -0.5 to 0.5 (implying that the numerator in Eq. (8) is equal to 1). These choices translate into a 95 percent probability that $\mu_{is,k}^j$ is in the range of $\boldsymbol{z}_i^j$.

The prior on the terminal node parameters handles overfitting with respect to setting S too large. Since S shows up in the denominator of Eq. (8), the terminal node parameters for a huge number of trees are increasingly forced to zero. Hence, if the number of trees increases each individual tree will contribute less to explaining the overall variation in the endogenous variables and the corresponding posterior variance of the function estimate will also decrease given that the prior becomes more informative in such a situation.

In our empirical work, we will use the same prior hyperparameters γ, α, and β for all equations in the VAR and for all $j \in \{f, g, u, v\}$. This choice reflects findings in Pratola, et al. (2020) that these hyperparameters also work well for hBART.

It is worth noting that this prior setup uses the actual data to scale the prior and is thus not a bona fide prior. In principle, it would be possible to not condition on the data and introduce additional scaling parameters. But given the excellent forecasting performance of BART across a range of applications using the priors stipulated in Chipman, George, and McCulloch (2010), we expect no substantive improvement in predictive performance and thus leave this option aside.

3.2 Full conditional posterior simulation

Posterior and predictive inference is done using MCMC methods. The full conditional posterior distributions of the model parameters are mostly available in closed form or can be obtained using a Metropolis-Hastings (MH) step. The conditional posteriors of the loadings $\boldsymbol{\Lambda}$, the factors $\boldsymbol{\delta}_t$, and the VAR coefficients $\boldsymbol{A}$ take well-known conditionally Gaussian forms and are thus discussed in the technical appendix. Here, we focus on how to sample the tree-specific structure used to approximate the unknown functions of the model.

Our MCMC algorithm exploits the fact that conditional on the factors and loadings, the equations of the model are independent (see also Kastner and Huber, 2020).[9] This implies that the model in Eqs. (1) and (2) can be written as a system of M independent regression models.

[9]For the fullBART specification, we estimate the model on an equation-by-equation basis conditional on the shocks in the previous equations.

The i^{th} equation closely resembles Eq. (6):

$$y_{it} = f_i(\boldsymbol{x}_t) + g_i(\boldsymbol{z}_t) + \boldsymbol{\lambda}_i \boldsymbol{\delta}_t + e_{it}, \quad e_{it} \sim \mathcal{N}\left(0, e^{v_j(\boldsymbol{w}_t)}\right), \tag{9}$$

which is a very general regression model with a scalar response. In what follows, we will discuss how to simulate the trees and terminal node parameters using Eq. (9), i.e., on an equation-by-equation basis.

3.2.1 Updating the trees

We sample the trees using the Bayesian backfitting strategy discussed in Chipman, George, and McCulloch (2010). This step samples each tree conditional on the remaining $S-1$ trees. Let $\tilde{\boldsymbol{z}}_{in}^f = \boldsymbol{z}_i^f - \sum_{n \neq s} l_{is}(\boldsymbol{X}|\mathcal{T}_{is}^f, \boldsymbol{\mu}_{is}^f)$, $\tilde{\boldsymbol{z}}_{in}^g = \boldsymbol{z}_i^g - \sum_{n \neq s} l_{is}(\boldsymbol{Z}|\mathcal{T}_{is}^g, \boldsymbol{\mu}_{is}^g)$, $\tilde{\boldsymbol{z}}_{in}^v = \boldsymbol{z}_i^v - \sum_{n \neq s} l_{is}(\boldsymbol{w}|\mathcal{T}_{is}^v, \boldsymbol{\mu}_{is}^v)$, and $\tilde{\boldsymbol{z}}_{kn}^u = \boldsymbol{z}_k^u - \sum_{n \neq s} l_{ks}(\boldsymbol{w}|\mathcal{T}_{ks}^u, \boldsymbol{\mu}_{ks}^u)$ denote partial residual vectors with the n^{th} tree l_{in} excluded.

Conditional on the partial residual vector $\tilde{\boldsymbol{z}}_{in}^j$ (for $j \in \{f, g\}$) and the full history of the latent error variances $\boldsymbol{h}_i = (v_i(\boldsymbol{w}_1), \ldots, v_i(\boldsymbol{w}_T))'$, which are modeled using hBART or SV, we simulate the tree structures $\mathcal{T}_{in}^j$ and terminal node parameters $\boldsymbol{\mu}_{is}^j$. Chipman, George, and McCulloch (2010) draw $\mathcal{T}_{in}^j$ marginally of $\boldsymbol{\mu}_{is}^j$:

$$p(\mathcal{T}_{in}^j|\tilde{\boldsymbol{z}}_{in}^j, \boldsymbol{h}_i) \propto p(\mathcal{T}_{in}^j) \int p(\tilde{\boldsymbol{z}}_{in}^j|\mathcal{T}_{in}^j, \boldsymbol{\mu}_{in}^j, \boldsymbol{h}_i)\ p(\boldsymbol{\mu}_{in}^j|\mathcal{T}_{in}^j, \boldsymbol{h}_i) d\boldsymbol{\mu}_{in}^j.$$

This integral can be solved analytically (up to a normalizing constant). To draw from $p(\mathcal{T}_{in}^j|\tilde{\boldsymbol{z}}_{in}^j, \boldsymbol{h}_i)$ we use the MH algorithm originally proposed in Chipman, George, and McCulloch (1998). Since the tree structure features a discrete state space, the MH algorithm specifies a transition kernel $q(\mathcal{T}_{in}^{j(a)}, \mathcal{T}_{in}^{j*})$ that is used to grow new trees $\mathcal{T}_{in}^{j*}$, conditional on the previously accepted tree structure ($\mathcal{T}_{in}^{j(a)}$), using one of four distinct moves with prespecified probabilities:

Grow The first possible move is to grow a terminal node. This move randomly selects a terminal node of $\mathcal{T}_{in}^{j(a)}$ and then proposes to split this terminal node into two new terminal nodes based on a random splitting rule. This move is selected with probability 0.25.

Prune The second move prunes a terminal node. It selects two terminal nodes and merges by collapsing the nodes below. This move is selected with probability 0.25.

Change This step randomly selects an interior node and changes the previously used splitting rule by assigning a new splitting rule. This splitting rule is obtained by randomly drawing a splitting variable from the prior (which follows a discrete uniform distribution) and a corresponding threshold. We select this move with probability 0.40.

Swap The final step swaps a splitting rule between parent and child nodes (a child node is one that arises from some other node). This move is used with probability 0.10.

These four moves yield a tree $\mathcal{T}_{in}^{j*}$ that is then accepted with probability:

$$\min\left(\frac{p(\mathcal{T}_{in}^{j*}|\tilde{\boldsymbol{z}}_{in}^{j},\boldsymbol{h}_i)}{p(\mathcal{T}_{in}^{j(a)}|\tilde{\boldsymbol{z}}_{in}^{j},\boldsymbol{h}_i)}\ \frac{q(\mathcal{T}_{in}^{j*},\mathcal{T}_{in}^{j(a)})}{q(\mathcal{T}_{in}^{j(a)},\mathcal{T}_{in}^{j*})},1\right). \tag{10}$$

This MH update has the advantage of being independent of the terminal node parameters and thus avoids issues with computationally involved reversible jump MCMC algorithms.

3.2.2 Updating the error variances

Sampling the trees used to approximate the functions that determine the conditional variance requires additional attention. Pratola, et al. (2020) propose a model that assumes that the trees enter the likelihood in product form. In this case, the same algorithm as the one used to approximate F and G can be used. However, if the number of trees is large, it commonly arises that one of the trees returns a volatility path that is very close to zero. In such a case, the original sampler of Pratola, et al. (2020) can get stuck. Our approach avoids this by first linearizing the model and then using an approximation to obtain a representation with Gaussian errors.[10]

To update $\mathcal{T}_{is}^{v}$ we first render the model conditionally Gaussian using the approximation proposed in Omori, et al. (2007).[11] Squaring and taking logs of the i^{th} element of $\boldsymbol{e}_t$, e_{it}, yields:

$$\log(e_{it}^2) = \sum_{is}^{S} l_i(\boldsymbol{w}_t|\mathcal{T}_{is}^{v},\boldsymbol{\mu}_{is}^{v}) + \varpi_{it}, \quad \varpi_{it} \sim \log\chi_1^2.$$

[10] For the conventional stochastic volatility specification, we use the same linear approximation and assume that the log-volatilities evolve according to independent AR(1) processes. The prior setup on the coefficients of the state equations mirrors the one proposed in Kastner and Frühwirth-Schnatter (2014); see also Appendix B.

[11] Here, we focus on the sampling step for $\mathcal{T}_{is}^{v}$. The sampling step for $\mathcal{T}_{is}^{u}$ is precisely the same with e_{it} replaced with δ_{jt}, the j^{th} element in $\boldsymbol{\delta}_t$.

ϖ_{it} is then simply approximated using a scale-location mixture of Gaussians with 10 components. The resulting model is a standard BART model with heteroskedasticity and a time-varying intercept. More precisely,

$$\log(e_{it}^2)|\xi_t = j \sim \mathcal{N}\left(\sum_{is}^{S} l_i(\boldsymbol{w}_t|\mathcal{T}_{is}^v, \boldsymbol{\mu}_{is}^v) + \mathfrak{m}_j, \mathfrak{s}_j^2\right), \tag{11}$$

with $\mathfrak{m}_j$ and $\mathfrak{s}_j^2$ being the mean and variance of the j^{th} Gaussian component, respectively. ξ_t denotes a component indicator that takes values between 1 and 10 with $\text{Prob}(\xi_t = j) = \mathfrak{q}_j$. The values of $\mathfrak{m}_j, \mathfrak{s}_j^2$, and $\mathfrak{q}_j$ are known and can be read off Table 1 in Omori, et al. (2007). Equation (11) is a standard BART model with time-varying intercept and variance, and the trees $\mathcal{T}_{is}^v$ can be sampled with the same MH step outlined above. The main difference with respect to the model outlined in Pratola, et al. (2020) is that they restrict the trees to be nonnegative and then, instead of assuming a sum, approximate the unknown positive function using a product of trees.

Conditional on the tree structures, the terminal node parameters for all of the different types of BART models we consider are easily simulated from independent Gaussian distributions. These take a standard form and resemble the one of a simple intercept model. The tree structure serves to allocate observations to different terminal nodes, and these observations are then consequently used to compute the posterior moments. If $\boldsymbol{y}_t$ contains severe outliers (such as the ones observed during the pandemic), BART will most likely group them together and the corresponding terminal node parameter will have a posterior variance that is equal to the inverse of the number of outliers plus the prior precision (which will be low; see Eq. (8)). Hence, the corresponding posterior variance will be large, which leads to wider predictive intervals and thus a higher probability of observing outliers under the posterior predictive distribution.

In case of the standard BART-based VAR, these methods are similar to the ones discussed in Huber and Rossini (2022) and a special case of the one developed in Huber, et al. (2020). The steps necessary to simulate each tree and the corresponding terminal node parameters individually are then combined with the steps outlined in Appendix B. This yields an MCMC algorithm that operates on an equation-by-equation basis by recursively sampling from the relevant full conditionals.

In large-scale VARs or regressions, the computational burden of MCMC methods can be a serious limitation to the implementation of Bayesian methods. The relevant algorithms (even in their most favorable implementations) result in a situation where the computational burden

increases enormously in M and p (and thus K). This has led papers such as Gefang, Koop, and Poon (2022) to use variational methods or other approximations. In the algorithms used with our BART models, this is not the case, since the number of explanatory variables has no direct impact on computation times because K only increases the space of possible decision rules that the algorithm needs to learn. This might cause mixing issues, but we have noticed that in cases where K is moderate to large (i.e., K up to 100), no mixing issues arise. Specifically, in our empirical work we take $30,000$ draws and discard the first $15,000$ draws as burn-in. MCMC convergence diagnostics based on the full sample corroborate findings in Chipman, George, and McCulloch (2010) and illustrate that our algorithm quickly converges toward the desired stationary distribution (see Table C.2 in the appendix).

In terms of estimation times our algorithm is fast. Depending on the BART variant, estimating the model with $M = 23$ equations and $T = 190$ observations takes between 0.5 to 1.5 hours on a 2020 Macbook Air M1. In light of the fact that larger values of K (i.e., more equations and/or more lags) do not imply a larger computational burden for the BART-part of the model, estimation of even larger models is feasible and can be carried out efficiently.

Appendix D contains a Monte Carlo study that presents additional evidence on the properties of our BART-based models.

4 Modeling and forecasting macroeconomic tail risks

After discussing data and evaluation metrics, we first assess the overall forecasting performance of the BART-based models, and then consider how well they perform when the focus is on the marginal predictive distributions of GDP growth, inflation in the GDP price index, and the unemployment rate. After showing that our different nonparametric models work well, we illustrate additional model features and qualitative properties of the predictive densities by focusing on the fullBART specification. Finally, we analyze the role of financial conditions in tail forecasting, and then focus on conditional forecasts during the episodes of the great financial crisis and of the COVID-19 pandemic.

4.1 Data overview, competitors and model specification

To assess the efficacy of BART-based models for macroeconomic forecasting, we evaluate the accuracy of real-time density and tail risk forecasts. We download our real-time data set from fred.stlouisfed.org. With real-time data vintages available beginning with 1996:Q4, our real-time

Table 1: Model overview.

Abbreviations	**Specification: F and G**	**Specification: $\boldsymbol{x}_t$, $\boldsymbol{z}_t$, $\boldsymbol{w}_t$**
Conditional mean models		
BVAR	F linear, G omitted	$\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$ $\boldsymbol{z}_t$ excluded
BART	F BART, G omitted	$\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$ $\boldsymbol{z}_t$ excluded
mixBART	F linear, G BART	$\boldsymbol{x}_t = \boldsymbol{z}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$
errorBART	F linear, G BART	$\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$ $\boldsymbol{z}_t = (\boldsymbol{\eta}'_{t-1}, \ldots, \boldsymbol{\eta}'_{t-p})'$
fullBART	f_i, g_i BART; $\boldsymbol{\Sigma}_t = \boldsymbol{H}_t$	$\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$ $\boldsymbol{z}_{it} = (\varepsilon_{1t}, \ldots, \varepsilon_{i-1,t})'$
Conditional variance models		
homosk	Homoskedastic model $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}$	
SV	Standard SV specification, u_i and v_j linear	$\boldsymbol{w}_t$: first lag of latent volatility
hBART	u_i, v_j are approximated with BART	$\boldsymbol{w}_t = (t, \boldsymbol{x}'_t)'$
Competing models		
TVP-VAR-SV	$F(\boldsymbol{x}_t) = \boldsymbol{A}_t\boldsymbol{x}_t$, G omitted Elements in $\boldsymbol{A}_t$ follow independent random walks	$\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$ $\boldsymbol{z}_t$ excluded
BQR	Univariate: $y_{it} = \boldsymbol{x}'_t\boldsymbol{\beta}_{ip} + \epsilon_{it}$, $\epsilon_{it} \sim \text{AL}_p(\sigma_{ip})$ Shocks follow an asymmetric Laplace (AL) distribution	$\boldsymbol{x}_t = (\boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$

Notes: Written in general form, the model is represented as $\boldsymbol{y}_t = F(\boldsymbol{x}_t) + G(\boldsymbol{z}_t) + \boldsymbol{\varepsilon}_t,$, $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{\Sigma}_t)$. The covariance matrix is decomposed as $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}\boldsymbol{\Omega}_t\boldsymbol{\Lambda}' + \boldsymbol{H}_t$ unless otherwise noted.

forecast sample begins with 1997:Q1 and ends with 2020:Q4.

The models are estimated with a set of 23 quarterly variables for the US. A number of studies have found that larger VARs of this dimension forecast as well as or better than smaller VARs (e.g., Banbura, Giannone, and Reichlin (2010); Koop (2013); and Carriero, Clark, and Marcellino (2019)). Our variable set is patterned after that of Giannone, Lenza, and Primiceri (2015). Note that we add to their variable set the broad national financial conditions index (NFCI) published by the Federal Reserve Bank of Chicago, which, starting with the work of Adrian, Boyarchenko, and Giannone (2019), is frequently used in the literature on assessing macroeconomic tail risks. With an eye to brevity, we focus our results on a few broad key indicators: GDP growth, inflation in the GDP price index, and the unemployment rate. Additional details on the data set are provided in Appendix A.

Our various BART models are compared to several other popular specifications. Most importantly, we consider various linear BVARs. All of our BVARs are specified (including prior choices where relevant) as special cases of the mixBART model with G removed and separate horseshoe priors used on the linear VAR coefficients and the factor loadings $\boldsymbol{\Lambda}$ (see also Appendix B). As a further competitor, we consider a small-scale TVP-VAR-SV[12] with factor stochastic volatility estimated using the non-centered parameterization with a horseshoe prior on the VAR

[12]TVP-VAR-SV uses GDP growth, inflation, the federal funds rate, unemployment, and the NFCI.

coefficients and the square root of the state-innovation variances (see also Frühwirth-Schnatter and Wagner, 2010; Huber, Koop, and Onorante, 2021). The priors on the FSV part of the TVP-VAR-SV are identical to those used for our BART models to enable direct comparisons. Moreover, in line with the preceding literature on assessing tail risk to macroeconomic variables, for each of the variables of interest we also include a Bayesian quantile regression (BQR; see Kozumi and Kobayashi, 2011) estimated for our full information set equipped with a horseshoe prior on the quantile-specific coefficients and a weakly informative prior for the scale parameters.[13] Further details on these models can be found in Appendix B. Table 1 provides a brief summary of all the models examined in the paper.

All the models we consider set $p = 5$, and the number of factors Q is set equal to the Ledermann bound, which implies a rather large number of factors; in our application with $M = 23$ variables, $Q = 16$.[14] Our shrinkage prior on the factor loadings, however, effectively prevents overfitting. In addition, we follow the general guidance of Chipman, George, and McCulloch (2010) and set the number of trees S for all our components equal to 250. As discussed in Section 2.2, as long as this number is not set too small, it does not impact forecasting accuracy significantly.

4.2 Forecast evaluation metrics

In evaluating real-time out-of-sample forecasts, we consider a range of metrics, many of which focus on tail risk.

As a baseline assessment of overall density accuracy, we use the continuous ranked probability score (CRPS) for marginal distributions (with equal weights for all quantiles of the predictive distribution) and the energy score (ES) for joint distributions. These metrics were developed in Gneiting and Raftery (2007). The CRPS, defined such that a lower number is a better score, is given by

$$\text{CRPS}_t(y_{it}) = \int_{-\infty}^{\infty} \left(\mathfrak{F}(z) - \mathbb{I}\{y_{it} \leq z\}\right)^2 dz = E_{\mathfrak{f}}|\hat{y}_{it} - y_{it}| - 0.5E_{\mathfrak{f}}|\hat{y}_{it} - \hat{y}'_{it}|,$$

where $\mathfrak{F}$ denotes the cumulative distribution function associated with the predictive density $\mathfrak{f}$, y_{it} $(1, \ldots, M)$ is the realization of the forecasted variable, $\mathbb{I}\{y_{it} \leq z\}$ is an indicator function

[13] Due to its univariate nature, direct-multi-step forecasts for the BQR model are used (as opposed to the iterative forecasts used with all other specifications).

[14] The Ledermann bound is the largest positive solution Q^* of the equation $(M - Q^*)^2 \geq M + Q^*$.

taking value 1 if $y_{it} \leq z$ and 0 otherwise, and $\hat{y}_{it}$ and $\hat{y}'_{it}$ are independent random draws from the posterior predictive density.

To assess joint forecast performance for the three main variables (GDP growth, inflation, and unemployment), we rely on the ES. The ES is a generalization of the CRPS, to which it collapses for $M = 1$:

$$\text{ES}_t(\boldsymbol{y}_t) = E_{\mathfrak{f}}||\hat{\boldsymbol{y}}_t - \boldsymbol{y}_t|| - 0.5E_{\mathfrak{f}}||\hat{\boldsymbol{y}}_t - \hat{\boldsymbol{y}}'_t||,$$

where $\hat{\boldsymbol{y}}_t$ and $\hat{\boldsymbol{y}}'_t$ are independent random vectors with distribution $\mathfrak{f}$. The ES provides a measure of overall density forecasting performance.

As a basic measure of accuracy of the lower tail risk forecast, we use the quantile score (QS), commonly associated with the tick loss function (see, e.g., Giacomini and Komunjer (2005)). The QS is computed as

$$\text{QS}_{\tau i,t} = (y_{it} - \mathcal{Q}_{\tau i,t}) \left(\tau - \mathbb{I}\{y_{it} \leq \mathcal{Q}_{\tau i,t}\}\right),$$

where $\mathcal{Q}_{\tau i,t}$ is the forecast quantile of the i^{th} variable at quantile τ, and the indicator function $\mathbb{I}\{y_t \leq \mathcal{Q}_{\tau i,t}\}$ has a value of 1 if the outcome is at or below the forecast quantile and 0 otherwise. We evaluate the QS using $\tau = 0.10$, 0.25, 0.75, and 0.90. We also evaluate tail forecast accuracy using two implementations of the quantile-weighted CRPS (qwCRPS) developed by Gneiting and Ranjan (2011) as a proper scoring function of the entire predictive density. The qwCRPS is computed as a weighted sum of quantile scores at a range of $J - 1$ quantiles:

$$\text{qwCRPS}_{it} = \frac{2}{J-1} \sum_{j=1}^{J-1} \omega(\tau_j) \text{QS}_{\tau_j i,t}, \tag{12}$$

with $\tau_j = j/J$. We rely on a grid of $J - 1 = 19$ quantiles $\tau \in \{0.05, 0.10, \ldots, 0.90, 0.95\}$ to compute these weighted scores. In one implementation (denoted qwCRPS-left), we set the weights to $\omega(\tau_j) = (1 - \tau_j)^2$ in order to target the left tail (downside risk), and in the other (denoted qwCRPS-right), we set the weights as $\omega(\tau_j) = \tau_j^2$ to target the right tail of the predictive distribution (upside risk).

The tables report averages of these score measures over the 1997–2020 period. We report scores relative to those of the benchmark BVAR-SV model. By all metrics, a ratio of less than 1 means that a given model is improving on the accuracy of the BVAR-SV baseline. To gauge statistical significance, we rely on Diebold and Mariano (1995) and West (1996) t-tests of

significance differences of scores, for each model compared to the benchmark.

To give some sense of performance over time, for each forecast metric we also report figures of mean scores computed recursively, relative to the recursive mean for the BVAR-SV benchmark. To be precise, the relative recursive mean for model i is $100(\text{FM}_{it}/\text{FM}_{\text{bench},t} - 1)$, where FM_{it} and $\text{FM}_{\text{bench},t}$ are the recursive averages of some forecast metric for model i and the benchmark model, respectively. For example, a value of -20 indicates that model i has a 20 percent better forecast performance according to the particular forecast metric used. The first two years of the holdout are not included, since we use the first eight observations to initialize the recursive means. Moreover, we consider the fluctuation test of Giacomini and Rossi (2010). This is a test that gauges out-of-sample forecast accuracy in potentially unstable environments. It tests for differences in forecast performance using a running window of forecast losses. We choose the running window to contain 20 percent of the holdout observations at a time and include the value of the test statistic and critical values for the 10 percent level of significance (positive values mean that we outperform the benchmark BVAR-SV).

We report results for horizons of $h \in \{1, 4, 8, 12\}$ quarters. Higher-order predictive densities are obtained through an iterative approach. The acronyms in the tables and figures can be understood by noting they combine the acronyms with the various specifications for the conditional mean (BVAR, BART, mixBART, errorBART, and fullBART) with the acronyms for different treatments of the conditional variance (SV and hBART). Results for homoskedastic versions of the models are labeled "homosk." in the tables.

4.3 Overall forecasting performance of the various BART models

In this subsection we consider the overall forecasting performance of the different BART-based models vis-á-vis the BVAR-SV. This comparison is based on the ES, which measures the joint density forecast performance for the three variables of interest. Subsequently, we will focus on the forecast performance for the variables individually.

Table 2 shows the average ES relative to the benchmark BVAR with SV. The relative ES values reported in Table 2 indicate that for $h = 1$ there are few gains from using BART relative to the BVAR-SV. Some BART specifications forecast slightly better than the BVAR-SV and others do slightly worse, but there are no substantial differences at this horizon. However, forecast gains become larger at longer forecast horizons and are, with few exceptions, statistically significant. These gains can be sizable, particularly for $h = 12$. At this horizon,

		h=1		h=4		h=8		h=12	
homosk.	BVAR	1.109	***	1.100	***	1.089	***	1.075	**
	BART	1.011		0.972	*	0.953	***	0.933	***
	mixBART	1.015		0.975	*	0.951	***	0.935	***
	errorBART	1.067		1.042	*	1.033		1.019	
	fullBART	1.035		0.965	**	0.932	***	0.917	***
SV	BVAR	2.194		2.321		2.508		2.658	
	BART	0.984		0.976	*	0.956	***	0.939	***
	mixBART	0.983		0.978	*	0.955	***	0.939	***
	errorBART	1.003		1.008		0.985		0.958	**
	fullBART	1.006		0.976		0.949	***	0.929	***
hBART	BVAR	1.002		0.985	**	0.990		0.984	**
	BART	0.981		0.972	**	0.946	***	0.933	***
	mixBART	0.983		0.970	**	0.948	***	0.933	***
	errorBART	1.060		1.019		0.999		0.974	**
	fullBART	1.007		0.970	**	0.926	***	0.908	***

	h=1	h=4	h=8	h=12
TVP−VAR−SV	0.977	1.004	0.980	0.947

Table 2: ES to measure joint forecast performance.

Notes: ESs are computed as the ratio with respect to the Bayesian VAR with SV. Asterisks indicate statistical significance of the Diebold-Mariano test for equal predictive performance at the 1, 5, and 10 percent level. The row associated with the benchmark (in grey) shows raw losses. Relative performance is illustrated with shades from red (benchmark better than alternative) to purple (alternative better than benchmark).

fullBART-hBART produces gains in forecast performance that can be as large as 10 percent. The fullBART-hBART specification is the generally best-performing model for $h = 8$ and 12, only losing to fullBART-homosk. at $h = 4$. For $h = 1$, the fullBART-hBART specification is very similar to all other approaches. Overall this leads us to select fullBART-hBART as our preferred model.

Comparing volatility specifications gives rise to a common finding in the literature: for linear VARs, it almost always pays off to allow for heteroskedasticity of some form (see, e.g., Clark (2011); Clark and Ravazzolo (2015)). The gains from adding SV or hBART to the linear BVAR are substantial, ranging from 7.5 percent (for $h = 12$) to 10 percent (for $h = 4$).

With the BART specifications, the effects of adding heteroskedasticity are smaller and the story is more nuanced. Consider, for instance, our best-performing fullBART model for $h = 12$. Relative to the homoskedastic variant, adding SV causes a very slight deterioration in forecast performance, whereas adding hBART causes a very slight improvement. Similar results are found for other BART specifications. The key point is that differences between homoskedastic and heteroskedastic variants are much less than they were with linear models. The nonparametric specification of the BART model evidently captures nonlinearities in the conditional mean process in such a way as to match or beat the accuracy gains that come from including the time-varying volatility of innovations in a linear VAR.

In other words, while other models perform poorly if the DGP is characterized by heteroskedasticity, BART is more flexible and, as we will show in Figure 5, produces very similar

		h=1	h=4	h=8	h=12
homosk.	BART	1.016 (−0.004)	0.994 (−0.022)	0.987 (−0.034)	0.972 (−0.039)
	mixBART	1.016 (−0.002)	0.995 (−0.020)	0.987 (−0.036)	0.972 (−0.037)
	errorBART	1.019 (0.048)	1.021 (0.021)	1.034 (−0.001)	1.045 (−0.026)
	fullBART	1.010 (0.025)	0.989 (−0.024)	0.985 (−0.053)	0.969 (−0.052)
SV	BART	1.005 (−0.021)	0.997 (−0.021)	0.988 (−0.033)	0.973 (−0.034)
	mixBART	1.003 (−0.021)	0.996 (−0.018)	0.988 (−0.034)	0.974 (−0.034)
	errorBART	0.995 (0.008)	1.000 (0.008)	0.999 (−0.014)	0.998 (−0.040)
	fullBART	1.024 (−0.019)	1.004 (−0.028)	0.995 (−0.046)	0.980 (−0.052)
hBART	BART	1.012 (−0.030)	0.992 (−0.020)	0.984 (−0.038)	0.968 (−0.035)
	mixBART	1.012 (−0.029)	0.991 (−0.020)	0.984 (−0.036)	0.969 (−0.036)
	errorBART	1.011 (0.050)	1.000 (0.019)	1.005 (−0.006)	1.008 (−0.034)
	fullBART	1.010 (−0.003)	0.989 (−0.019)	0.982 (−0.057)	0.967 (−0.058)

Table 3: ES for forecasts normalized to BVAR-SV point forecasts, relative to BVAR-SV.

Notes: Predictive densities for all models are normalized such that the point forecasts coincide with those of the benchmark. Consequently, any differences are driven solely by higher-order moments of the predictive distribution. Values shown in parentheses are differences of the relative ES for original densities minus those for normalized densities.

predictive densities irrespective of the specification for the conditional variance. This is because BART is doing so well at fitting the dynamics via the flexible specification for the conditional mean that there is little role left for the conditional variance. Evidence from our Monte Carlo study (see Figures D.1 and D.2 in Appendix D) backs this claim. In particular, it shows that a homoskedastic BART specification accurately recovers the true mean function in the presence of heteroskedastic shocks. By contrast, linear models that ignore heteroskedasticity perform poorly in this situation and this partly explains the weak forecasting performance of the BVAR with homoskedastic shocks.

BART may be leading to improved energy scores either because of an improvement in the point forecasts or because of an improvement in other aspects of the predictive distribution. To investigate which aspect is of most importance, we re-calculated the energy scores in Table 2 using predictive distributions that have been re-centered to the mean forecasts of the BVAR-SV. Thus, the forecasts for all models will have the same mean and any difference in energy scores is due to differences in the modelling of the rest of the predictive distribution, including its tails. To be precise, for each MCMC simulation of a draw from the predictive distribution of one of our BART models, we added the difference between the predictive mean of the BART model and the predictive mean of the BVAR-SV (both predictive means averaged over all draws).

The results of this analysis are provided in Table 3. This table provides the ESs based on the normalized predictive densities (relative to BVAR-SV). The differences (in percentage points) to the relative ESs reported in Table 2 are shown in parenthesis. Negative values indicate that the normalized densities yield higher ESs, whereas positive differences imply that normalizing yields smaller ESs.

The table suggests that these re-centered predictive distributions (with a few exceptions, particularly for $h = 1$) do have lower ESs than those of the BVAR-SV. But most of these ESs are higher than the non-normalized ones. This indicates that gains arise from both better point forecasts and other aspects of the predictive distribution. These differences increase with the forecast horizon, implying that by conditioning on the forecast mean of the linear model we lose important information for higher-order predictions.

To understand how much of the accuracy gains arise from more accurate point forecasts, we can compare the percentage point differences to the level of the original relative ES. For instance, in the case of fullBART-hBART and $h = 12$, improvements in energy scores are close to 10 percent if the non-normalized predictive density is considered (see Table 2). These gains decrease by almost 6 percentage points if we consider the normalized predictive distribution. One interpretation of this would be that whereas higher-order features of the predictive distribution explain around 4 percentage points of the overall gain (in terms of non-normalized densities), better point forecasts dominate the overall gains, contributing around 6 percentage points. The finding that nonparametric techniques offer gains in terms of point forecasts is consistent with recent findings of Medeiros, et al. (2021).

This analysis can be repeated for all other models under consideration, revealing a remarkably consistent feature. If a BART-based model improves upon the VAR, most of these improvements are driven by more precise point forecasts but the larger flexibility of BART allows for predictive distributions that display heavy tails, skewness, and richer covariance structures across shocks. All of these features are key during turbulent times but also pay off in normal periods.

To analyze whether forecast performance changes over time, Figure 1 reports, for $h = 1$ and $h = 4$, the relative recursive mean of ES (top panel) and the Giacomini and Rossi (2010) fluctuation test (bottom panel) between fullBART-hBART and the BVAR-SV.[15] The relative recursive mean score ratios in the top panel show that the full sample performance of the models largely holds up over time. For horizons of $h = 4, 8, 12$, the fullBART-hBART consistently beats the BVAR-SV benchmark over time, whereas for $h = 1$, the benchmark has the advantage for most of the recursive sample, except in the early years and for some deterioration late in the sample, with the pandemic. The fluctuation test results in the bottom panel indicate that the fullBART-hBART's advantages are statistically significant at the horizons of $h = 8, 12$, whereas

[15] Plots for other models look very similar and are available upon request from the corresponding author.

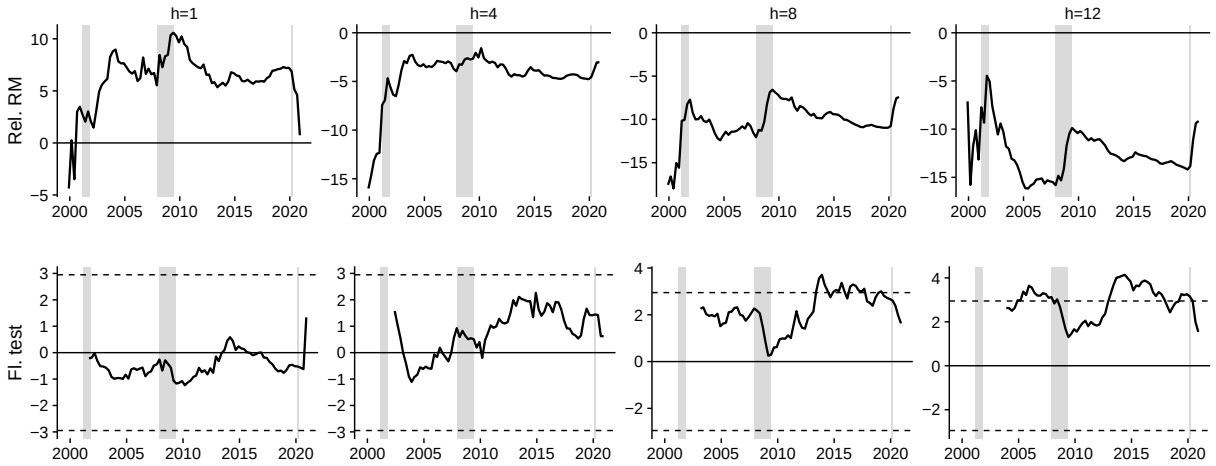

Figure 1: Relative recursive mean and fluctuation test statistic for joint forecast performance.

Notes: Joint forecast performance is measured by the ES. The chart compares fullBART-hBART as the on average best-performing model to the benchmark BVAR-SV. For details on the fluctuation test, see Giacomini and Rossi (2010). Dashed lines indicate critical values for a 10 percent level of statistical significance; the solid horizontal line marks zero. Negative values of relative recursive mean and positive values of the fluctuation test imply that the fullBART-hBART model outperforms the benchmark.

for $h = 1, 4$, the test does not reject the null of equal accuracy over time.

4.4 Variable-specific tail forecasting performance

The ES is a measure of overall density forecasting performance. To investigate whether using BART pays off for our three focus variables, we now consider the different variants of the CRPS across variable types. For brevity we focus on the three best-performing BART models (in terms of ES, averaged over the forecast horizons). In this exercise, we also include other benchmark models commonly used in the forecasting literature.

Table 4 provides results on accuracy measures that refer to the entire predictive density, including the CRPS, qwCRPS-right, and qwCRPS-left.

In the case of GDP growth, the BART specifications offer consistent improvements in CRPS, small to modest (roughly 3 percent for $h = 12$) in the left tail and a little larger in the right tail (roughly 5 percent for $h = 12$). Among the BART specifications included in the table, fullBART-hBART tends to be a little better than the others, but the differences are small. On balance, the performance of the BART specifications may be seen as comparable to that of the BQR model. In a number of cases (combinations of metrics and horizons), the BART and BQR score ratios are similar, whereas, in others, BQR fares either a little or modestly better than BART (e.g., qwCRPS-right with $h = 12$) or worse (e.g., qwCRPS-left with $h = 12$). The TVP-VAR-SV model for the three variables is beaten by the benchmark, except at the $h = 1$ horizon.

		CRPS				qwCRPS−left				qwCRPS−right			
		h=1	h=4	h=8	h=12	h=1	h=4	h=8	h=12	h=1	h=4	h=8	h=12
GDPC1	BVAR−SV	1.965	2.002	2.075	2.133	0.607	0.613	0.632	0.656	0.539	0.557	0.586	0.599
	TVP−VAR−SV	0.966	1.023	1.035	1.026	0.972	1.026	1.059	1.031	0.967	1.022	1.001	1.009
	BQR	0.953*	0.989	0.973	0.929	1.014	1.109	1.084	1.052	1.011	0.993	0.972	0.904***
	BART−hBART	0.983	0.995	0.982	0.979	0.991	1.006	1.005	0.987	0.979	0.987	0.954**	0.963
	mixBART−hBART	0.984	0.993	0.985	0.979	0.990	1.003	1.008	0.986	0.983	0.986	0.955**	0.963
	fullBART−hBART	0.993	0.992	0.974	0.972	1.005	1.009	1.003	0.984	0.984	0.974	0.934***	0.945*
GDPCTPI	BVAR−SV	0.565	0.703	0.892	1.042	0.176	0.207	0.239	0.267	0.157	0.207	0.285	0.346
	TVP−VAR−SV	1.063	0.931	0.781*	0.710**	1.047	0.945	0.868	0.857*	1.074	0.906	0.712*	0.600***
	BQR	1.164	1.033	0.964	0.911	1.054	1.052	1.157	1.235	1.435*	1.089	0.896	0.733***
	BART−hBART	0.990	0.863***	0.822***	0.797***	0.970	0.865***	0.835***	0.838***	1.013	0.861***	0.808***	0.761***
	mixBART−hBART	0.996	0.867***	0.822***	0.798***	0.977	0.872***	0.839***	0.839***	1.015	0.860***	0.806***	0.762***
	fullBART−hBART	1.091	0.864***	0.757***	0.708***	1.042	0.921	0.836***	0.811***	1.137*	0.800***	0.688***	0.623***
UNRATE	BVAR−SV	0.267	0.304	0.321	0.332	0.068	0.079	0.083	0.086	0.088	0.097	0.104	0.107
	TVP−VAR−SV	0.973	0.981	1.046	1.102*	1.003	1.026	1.100	1.153*	0.949*	0.946	1.003	1.057
	BQR	1.154	0.974	0.971	0.973	1.527	1.036	1.028	1.012	1.020	1.054	1.061	1.079
	BART−hBART	0.948	0.984	1.015	1.036**	0.962	1.005	1.019	1.033	0.930	0.968	1.011	1.033***
	mixBART−hBART	0.951	0.986	1.018	1.035**	0.963	1.006	1.021	1.034	0.934	0.970	1.013	1.032***
	fullBART−hBART	0.964	0.978	1.008	1.035***	0.980	0.995	1.009	1.031	0.940	0.960**	1.003	1.032***

Table 4: Variants of CRPS by variable.

Notes: Relative performance is illustrated with shades from red (benchmark better than alternative) to purple (alternative better than benchmark).

Our proposed BART models yield the biggest benefits to inflation forecasting. While BART's performance at the $h = 1$ horizon is comparable to that of the benchmark, it is materially better at longer horizons, increasingly so as the horizon rises. The benefits of the BART models are moderately greater in the right tail than the left but sizable in both. As examples, the fullBART-hBART model's advantage over the benchmark increases from about 12 percent for $h = 4$ to 20 percent for $h = 12$ using the qwCRPS-left metric, and it rises from about 22 percent for $h = 4$ to 32 percent for $h = 12$ using the qwCRPS-right metric. BART is also consistently at least as good as — better in most cases — BQR. The same applies for BART as compared to TVP-VAR-SV, although the TVP-VAR-SV specification usually beats BQR.

By the CRPS metrics, BART's forecast performance is more mixed for the unemployment rate than for GDP growth and inflation. The BART specifications yield modest improvements in forecast accuracy for horizons of $h = 1, 4$, slightly more in the right tail than the left. At the longer horizons, the BART models are about as accurate as or slightly less accurate than the BVAR-SV benchmark. It remains the case, however, that the BART specifications perform at least as well as the BQR and TVP-VAR-SV alternatives.

Figure 2 is similar to Figure 1 and depicts the recursive mean of the qw-CRPSs of fullBART-hBART relative to the BVAR-SV over time.[16] To gauge whether changes in the relative performance for a given point in the hold-out are statistically significant, the lower panels of each figure show the evolution of the fluctuation test statistic over time. For readability,

[16] Plots that show the relative scores for all other models and volatility specifications are available in the online appendix.

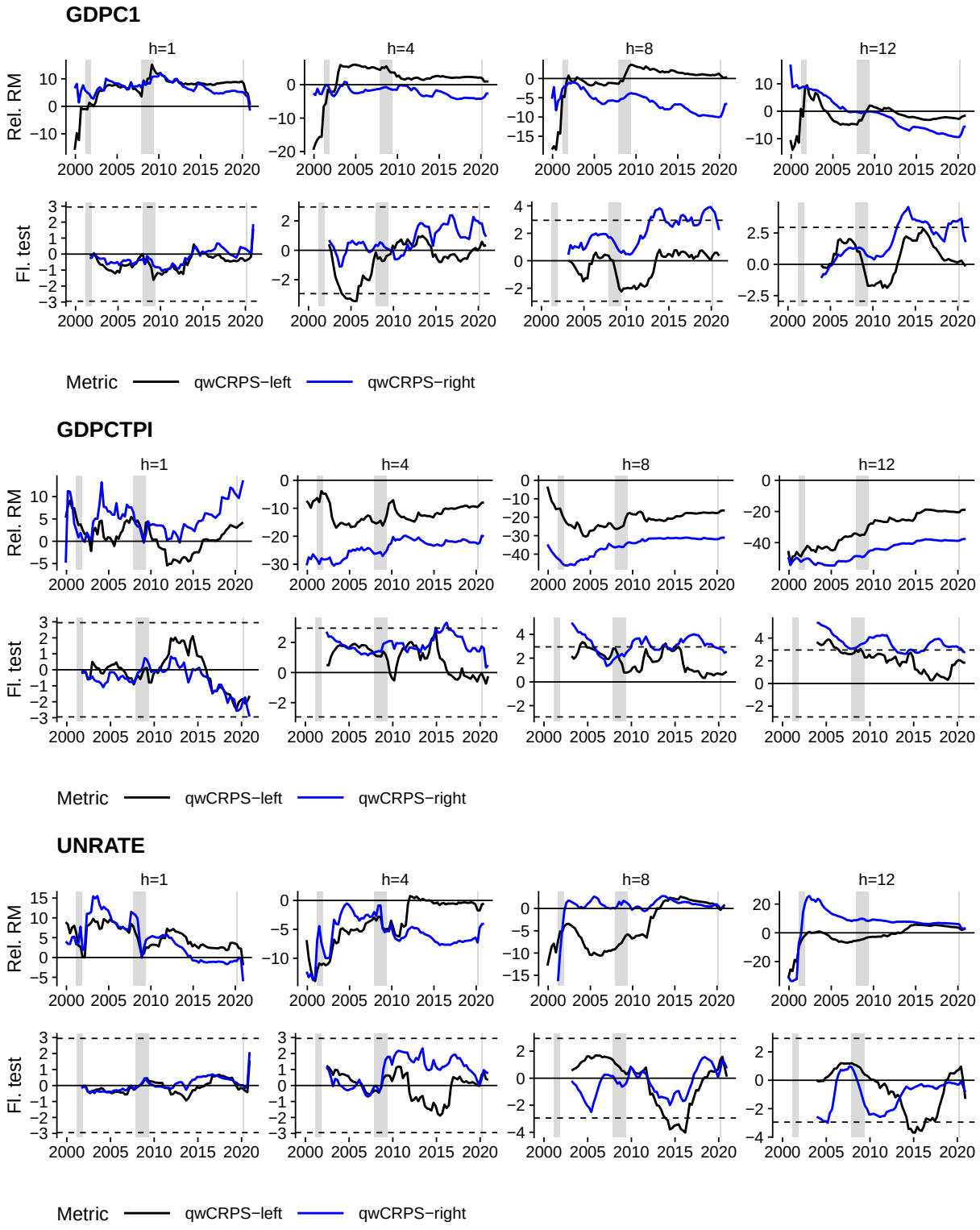

Figure 2: Quantile weighted CRPSs; relative recursive mean and fluctuation test for fullBART-hBART relative to BVAR-SV.

the figures report results using the qwCRPS-left and qwCRPS-right metrics.

Broadly, the relative scores given in the top row of each panel indicate that, in many cases, the forecast performance of our preferred fullBART-hBART specification relative to the BVAR-SV benchmark is largely stable over time, although with some instabilities, more so for

		QS10				QS25				QS75				QS90			
		h=1	h=4	h=8	h=12	h=1	h=4	h=8	h=12	h=1	h=4	h=8	h=12	h=1	h=4	h=8	h=12
GDPC1	BVAR−SV	0.87	0.89	0.95	1.01	1.14	1.15	1.17	1.24	1.00	1.04	1.11	1.14	0.63	0.70	0.76	0.80
	TVP−VAR−SV	0.99	1.03	1.03	1.01	0.97	1.02	1.06	1.00	0.95	1.02	0.97	0.99	1.02	1.01	0.91**	0.95**
	BQR	1.07	1.28	1.14	1.15	1.01	1.09	1.09	1.03	0.96*	0.95	0.93**	0.84***	1.21	1.04	0.96	0.86**
	BART−hBART	1.01	1.01	0.99	0.95***	1.00	1.01	1.02	0.98	0.97	0.99	0.93**	0.94*	1.03	0.97	0.92***	0.92**
	mixBART−hBART	1.01	1.01	1.00	0.95***	1.00	1.00	1.03*	0.98	0.98	0.99	0.93**	0.94*	1.03	0.97	0.92***	0.92*
	fullBART−hBART	1.02	1.02	0.99	0.96	1.02	1.02	1.02	0.97	0.96	0.97	0.90***	0.92**	1.03	0.94*	0.89***	0.89**
GDPCTPI	BVAR−SV	0.21	0.23	0.23	0.28	0.35	0.40	0.43	0.44	0.29	0.40	0.57	0.71	0.17	0.24	0.34	0.43
	TVP−VAR−SV	0.98	0.92	1.03	1.03	1.05	0.93	0.92	1.02	1.09	0.91	0.70*	0.55***	1.00	0.74***	0.62**	0.53***
	BQR	1.11	1.22	1.71	1.89*	1.00	0.96	1.20	1.37*	1.39*	1.05	0.82	0.66***	2.43	1.21	0.98	0.59***
	BART−hBART	0.95	0.87**	0.85	0.82*	0.96	0.86**	0.85***	0.90	1.01	0.85***	0.81***	0.73***	1.06	0.86***	0.78***	0.72***
	mixBART−hBART	0.95	0.88**	0.84	0.83*	0.98	0.88**	0.86***	0.89	1.01	0.85***	0.80***	0.74***	1.06	0.86***	0.78***	0.72***
	fullBART−hBART	1.01	1.01	0.91	0.93	1.01	0.92	0.89**	0.92***	1.18**	0.76***	0.66***	0.59***	1.18	0.69***	0.60***	0.54***
UNRATE	BVAR−SV	0.08	0.10	0.11	0.11	0.13	0.14	0.15	0.15	0.16	0.18	0.20	0.20	0.14	0.16	0.17	0.18
	TVP−VAR−SV	0.98	1.08	1.10	1.13	1.04	1.04	1.14	1.18*	0.93**	0.92**	0.98	1.03	0.94**	0.92***	0.96	0.99
	BQR	2.20	1.03	1.06	1.01	1.61	1.05	1.00	1.00	0.93*	1.03	1.02	1.05	1.01	1.10	1.15	1.18
	BART−hBART	1.03	1.02	0.99	0.97	0.96	1.01	1.02	1.03	0.92	0.96	1.01	1.03***	0.93	0.94**	0.99	1.00
	mixBART−hBART	1.01	1.02	0.99	0.97	0.96	1.01	1.02	1.03	0.93	0.97	1.01	1.03***	0.94	0.95**	0.99	1.00
	fullBART−hBART	1.06	1.00	0.98	0.97	0.96	1.00	1.01	1.03	0.93	0.95**	0.99	1.02	0.94	0.93***	0.99	1.01

Table 5: Quantile scores by variable type.

Notes: Relative performance is illustrated with shades from red (benchmark better than alternative) to purple (alternative better than benchmark).

GDP growth and unemployment than for inflation. In the case of inflation, there appears to be modest instability for $h = 1$, but the fluctuation test reported in the lower panel does not reject the null of equal accuracy over time. BART's accuracy gains are largely stable over time at the longer forecast horizons, with significant rejections of the null, more frequently in the right tail than the left (in keeping with the full-sample result noted above of modestly larger accuracy gains in the right tail than the left). For GDP growth, the relative accuracy measures show stability over time in some cases (e.g., qwCRPS-right for $h = 8$ or less) and some instability in others (e.g., qwCRPS-right for $h = 12$). The null of stability over time is only rejected for some periods of time using qwCRPS-right for $h = 8, 12$ and using qwCRPS-left for $h = 4$. Scores for the unemployment rate show modestly more instability, but the null of equal accuracy over time is only rejected in a couple of instances, such as in the left tail for $h = 8, 12$.

To shed additional light on performance gains arising from different parts of the predictive distribution, we examine the tail forecasting performance by means of quantile scores. Table 5 has a format similar to that of Table 4, but presents QS results instead of CRPS results.

Broadly speaking, the QS results are quite similar to the CRPS results. Once again, BART yields the largest payoff in forecasting tail risks to inflation, with gains that rise with the forecast horizon and are greater in the right tail than the left. In this case, fullBART-hBART often offers accuracy gains greater than those of the other BART specifications. The fullBART-hBART's accuracy gains over the BVAR-SV benchmark reach more than 40 percent for the 75 and 90 percent QS at $h = 12$.

In the case of GDP growth, BART also offers some gains in tail risk forecast accuracy as

measured by the QS, modestly more so in the left tail than the right, and more so at longer horizons than shorter. BART's forecast performance is more mixed for the unemployment rate, offering modest gains at shorter horizon forecasts of the right tail and generally matching the accuracy of the benchmark BVAR-SV model. Throughout these results, the BART models forecast at least as well as the BQR and VAR-TVP-SV alternatives. The BART forecasts achieve statistically significant gains in more cases (across horizons and quantiles) than do these alternative models, particularly for GDP growth and inflation and less so for the unemployment rate.

Turning to the question of whether this strong performance of the different BART approaches holds throughout the hold-out period or is specific to certain time periods, Figure 3 displays recursive averages of the quantile scores for $\tau \in \{0.1, 0.9\}$. The patterns in these quantile scores are broadly similar to those for the qwCRPS-left and qwCRPS-right measures discussed above. Focusing on the fluctuation test results, the null of stability over time is rarely rejected for the unemployment rate, but rejected in a good part of the sample for the right-tail forecasts of GDP growth and inflation at longer horizons.

4.5 A deeper examination of fullBART

In the previous subsections we have shown that fullBART yields forecasts that are highly competitive and, according to the ES, the most precise ones over all forecast horizons. To shed light on what features of the predictive density drive the good forecasting performance, we now analyze in more detail several characteristics of the fullBART specification.

4.5.1 What variables drive conditional mean and variance dynamics?

Before focusing on the properties of the predictive distributions of fullBART we investigate how our preferred fullBART-hBART specification extracts information from the full sample to model US macroeconomic dynamics. Figure 4 shows percentages of how often fullBART decides to use each variable in its splitting rules per equation in relation to the total number of splits (summing over the respective lags). Assessing splitting rules to determine the importance of a variable is less effective when S is large, since the abundance of trees leads to mixing many irrelevant predictors with relevant ones; see also the discussion in Chipman, George, and McCulloch (2010, p. 276). Nonetheless, we find that some noteworthy patterns emerge.

It turns out that all variables matter. This holds for the conditional mean and the con-

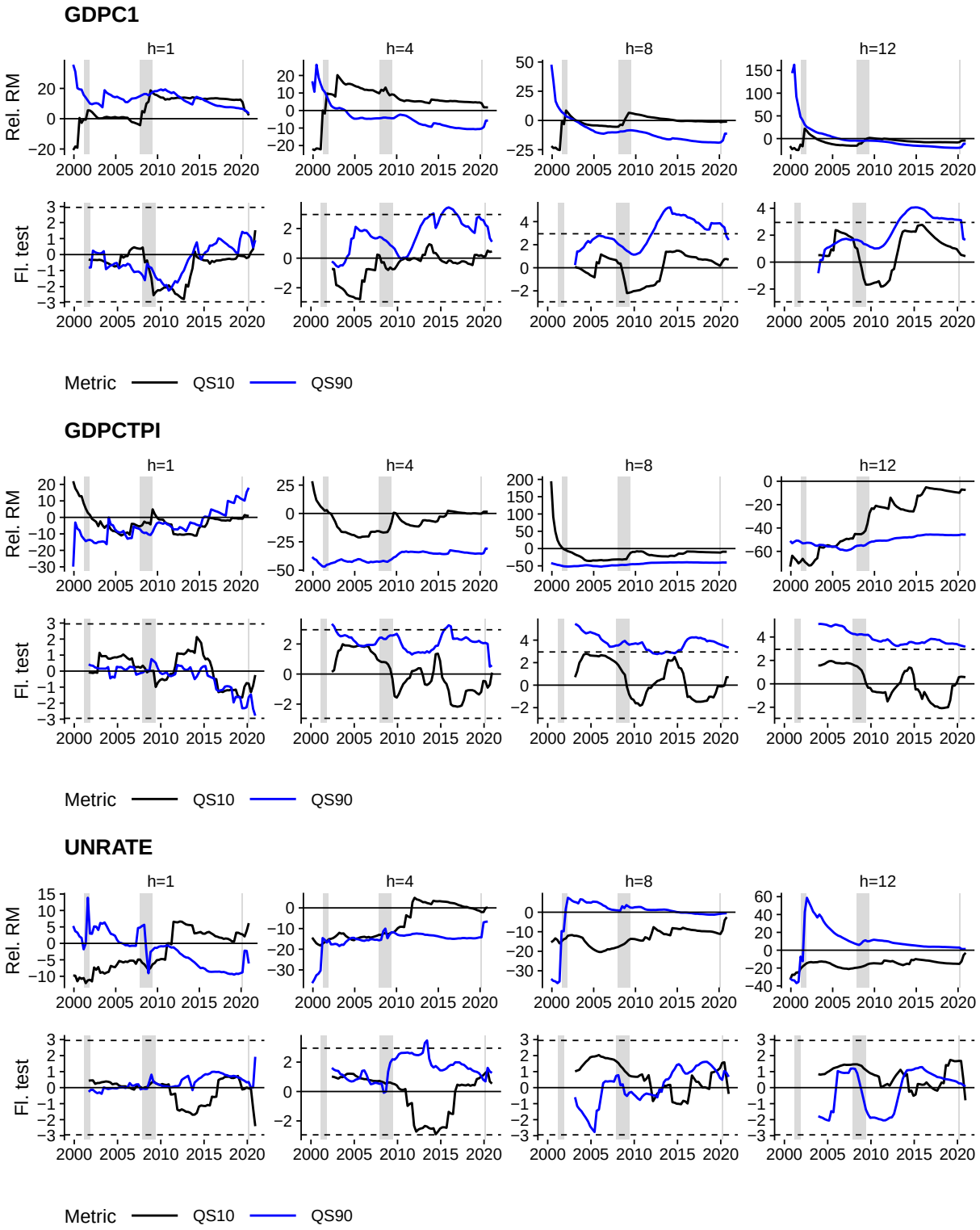

Figure 3: Quantile scores; relative recursive mean and fluctuation test for fullBART-hBART relative to BVAR-SV.

ditional variance. This result is more pronounced for the conditional variance. The inclusion percentages in the hBART component of the model suggest a dense volatility model with less heterogeneity across the three equations than is the case in the conditional mean part of the model.

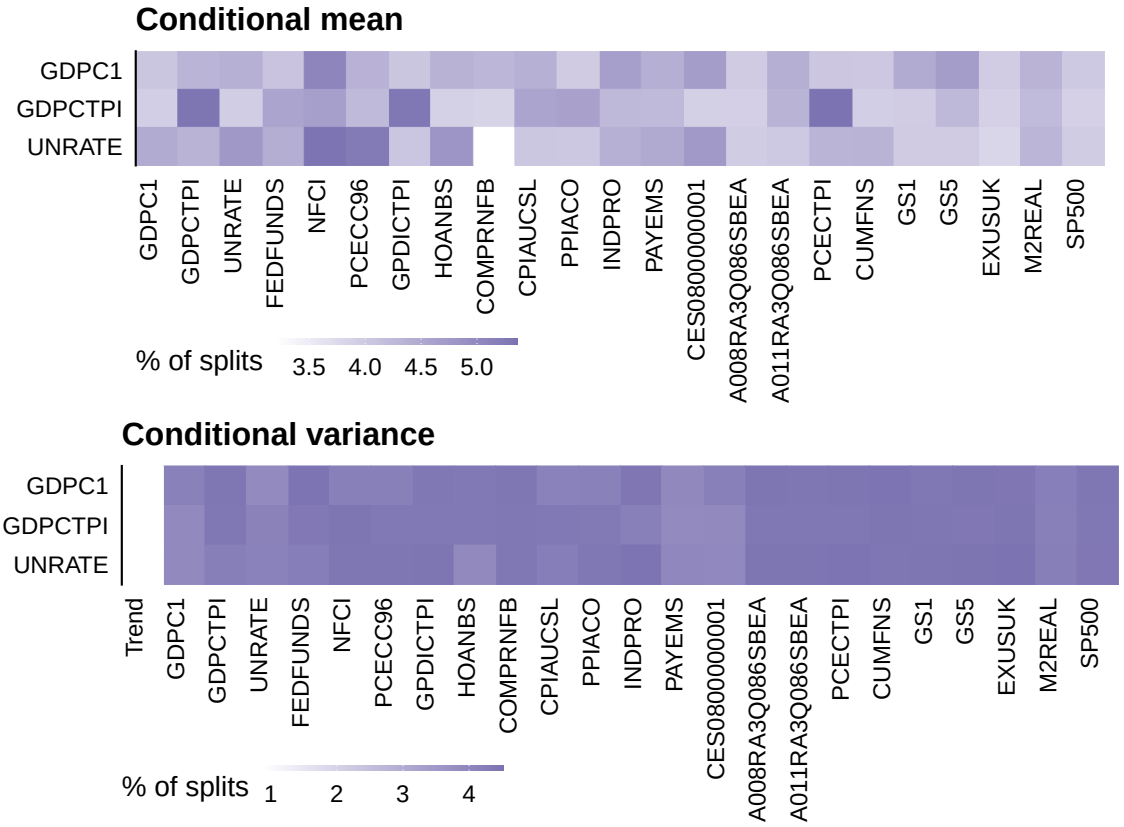

Figure 4: Percentages of variables in splitting rules relative to the total number of splits.

The conditional mean part suggests that some variables seem to be more important than others. For instance, we find that the NFCI plays an important role for GDP growth and unemployment, while it appears to be (slightly) less relevant for inflation. By contrast, for inflation we observe that lagged inflation and price measures (such as GPDICTPI and PCECTPI) frequently show up in the conditional mean.

To investigate whether controlling for heteroskedasticity changes mean relations, Figure C.2 in the appendix reproduces these heatmaps for all three volatility models we propose. The key finding is that, across our three focus variables, the same set of variables frequently shows up. This indicates that the conditional mean estimates are informed by similar covariates and these do not change if we use more flexible models to capture conditional variance dynamics.

These heatmaps do not convey information on how complicated the individual trees are. They suggest that different variables shape our estimates of the conditional mean and conditional variance. But this can be achieved in two ways using BART. First, it could arise through a few rather complicated trees that feature many of the different variables as splitting variables. Second, it could arise through many simple trees that only feature very few splitting variables but these differ across trees. Table C.1 in the appendix provides a few summary statistics for

the different trees. In this table we find that the trees are indeed rather simple (consistent with our prior) and do not differ across equations and volatility specifications, providing evidence that if a given variable shows up in a splitting rule, this typically happens through a simpler tree.

4.5.2 Properties of predictive distributions

We now consider the predictive distributions of fullBART and compare them to the ones obtained from the linear BVAR. Figure 5 provides, for the three variables, time series of the 5/95 percent and 10/90 percent quantiles of the predictive distributions from the BVAR and fullBART models for different volatility specifications. So that the extreme volatility of 2020 induced by the COVID-19 pandemic does not obscure scales, the charts have separate panels for the 1997–2019 and 2020 periods. In the interest of chart readability, the figures provide results for just the 1- and 4-quarter-ahead horizons; results for the 8- and 12-quarter-ahead horizons are very similar to those for the 4-quarter-ahead horizon.

This figure tells a clear story about why we do not gain much from controlling for heteroskedasticity if we use a BART-based specification. In the upper panel (i.e., for the BVAR) we find that forecast intervals differ sharply across volatility specifications. The BVAR with homoskedasticity produces much wider intervals than the ones from the BVARs with either SV or hBART. This is driven by the fact that the homoskedastic model mixes over low, intermediate, and high volatility periods and thus produces predictive intervals that are either too wide (in a low volatility regime) or too narrow (in a high volatility regime).

By contrast, and this is perhaps the most striking feature of the BART-based predictive distribution, across volatility specifications there is strong similarity of the forecast intervals for fullBART, for both forecast horizons. Irrespective of the volatility specification, forecast intervals only change slightly for GDP growth, while they are almost identical for inflation and unemployment. This, again, corroborates our finding above that BART is capable of controlling for model mis-specification with respect to assumptions on the error variances.

Considering the dynamics and magnitudes of the corresponding volatility estimates corroborates this finding. For brevity these are provided in the appendix (see subsection C.1) and here we summarize the main findings. In the case of the BVARs we find that volatility estimates strongly differ in magnitudes, with the homoskedastic specification providing the largest estimate of the error variances. For fullBART, there are differences in terms of both the shape and

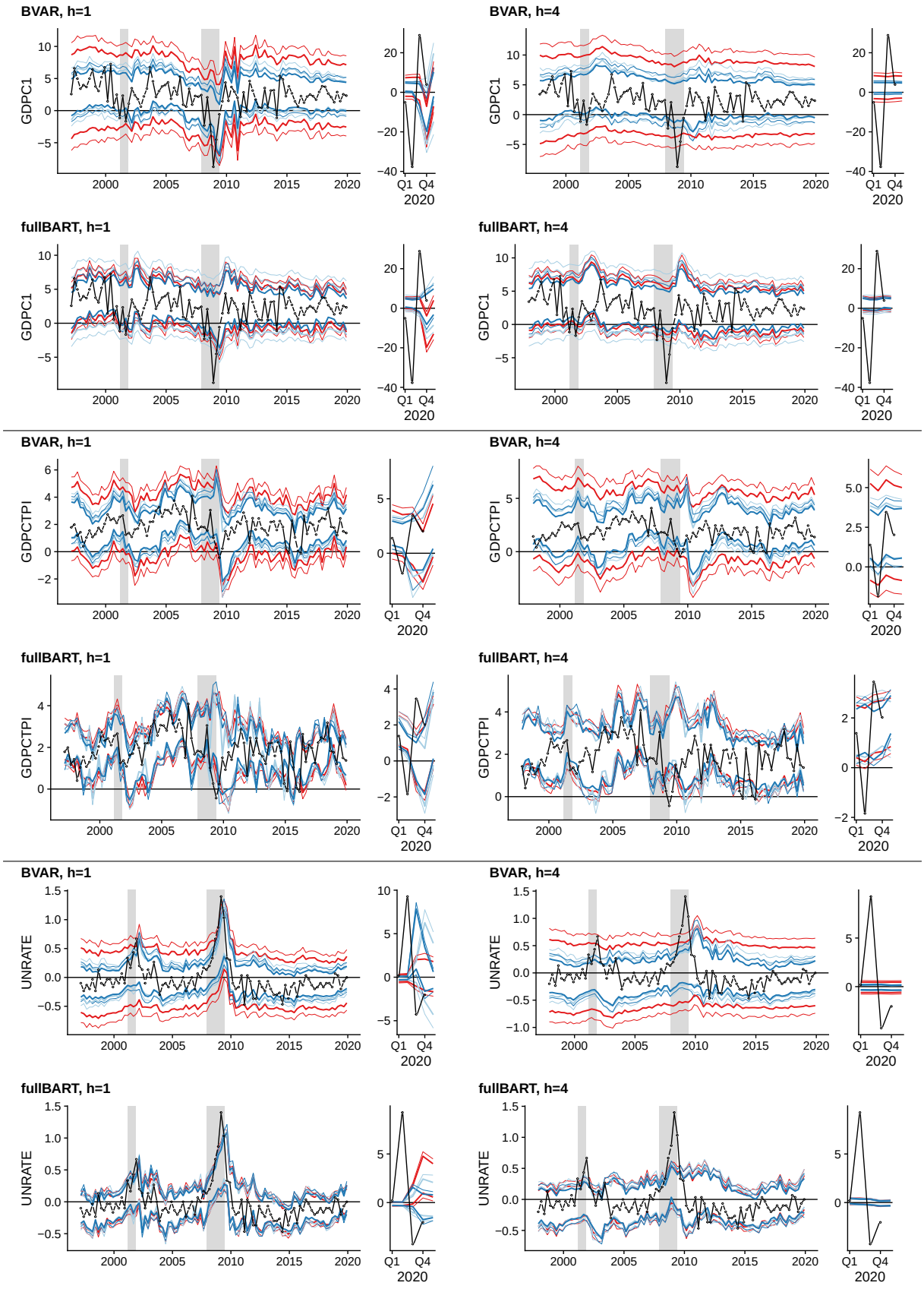

Figure 5: Predictive densities: BVAR versus fullBART.

Notes: Constant volatility (—), SV (—), hBART (—). Thin colored lines mark the 5/95$^{\text{th}}$ percentile, thick lines the 10/90$^{\text{th}}$ percentile. Black lines and points are realizations of the final vintage.

the magnitude of the different volatility processes but the magnitudes are substantially smaller. The immediate consequence of this finding is that the predictive distribution is mostly driven by our flexible modeling of the conditional mean, supporting the idea that most of the benefits of the nonparametric approach are obtained in its modeling of the conditional mean as opposed to the conditional variance.

4.5.3 Nonlinear features in predictive densities

Visual inspection of the predictive densities masks possible nonlinear features. To assess if and when predictive distributions depart from linearity and normality, we rely on a linear approximation of the BART model to obtain a linear VAR representation. Similar approximations have been proposed in Crawford, et al. (2018) and adopted in Huber, et al. (2020).

All results shown up to this point are based on the exact predictive distribution that is available through simulation. In this section, we compare these exact predictive distributions to the ones obtained from a linear approximation. The approximation linearly projects $\boldsymbol{\Xi} = (\boldsymbol{X}, \boldsymbol{Z})$ on a $T \times M$ matrix of nonparametric functions, $\boldsymbol{F}$, with typical t^{th} row $F(\boldsymbol{x}_t) + G(\boldsymbol{z}_t)$:

$$\tilde{\boldsymbol{A}} = \text{Proj}(\boldsymbol{\Xi}, \boldsymbol{F}) = \boldsymbol{\Xi}^{\dagger}\boldsymbol{F},$$

with $\boldsymbol{\Xi}^{\dagger}$ being the Moore-Penrose inverse of the matrix of explanatory variables, $\boldsymbol{\Xi}$. The reader is referred to the discussion in Huber, et al. (2020) to justify this approximation. In essence, it can be shown that, at the T observations, $\boldsymbol{F} \approx \boldsymbol{X}\tilde{\boldsymbol{A}}$.

After having obtained $\tilde{\boldsymbol{A}}$ we can iteratively compute multi-step forecasts using standard formulas for forecasting in VAR models. This yields an approximate predictive distribution $\hat{p}(\boldsymbol{y}_{t+h}|\boldsymbol{y}_t)$. In principle, if the DGP is linear we would expect that $\hat{p}(\boldsymbol{y}_{t+h}|\boldsymbol{y}_t) \approx p(\boldsymbol{y}_{t+h}|\boldsymbol{y}_t)$. Hence, the distance between the approximate and the exact predictive distribution yields insights into the extent of nonlinearities. To formalize the idea of distance between distributions, we use the Kullback-Leibler divergence (KLD). The KLD will serve as a measure of the importance of nonlinearities. If we observe substantial divergence between the two predictive densities, this indicates that exact distributions feature substantial nonlinearities (since approximation errors become comparatively large).

Figure 6 shows the measure for the fullBART-hBART specification over time and at different forecast horizons as a heatmap. The time axis here refers to the date when the forecast was made.

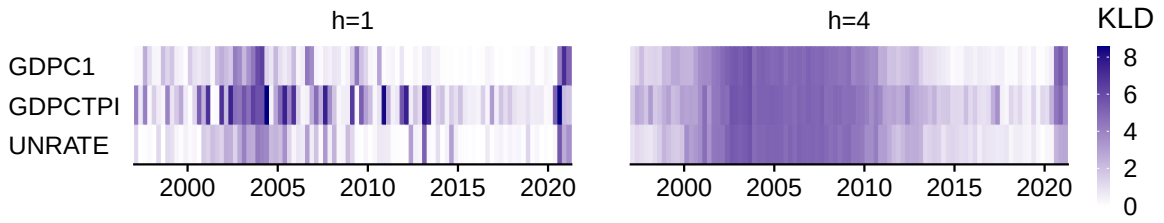

Figure 6: KLD between exact and approximate predictive distribution for fullBART-hBART.

A common pattern found for all three of the variables is that the KLD becomes sizable during the global financial crisis and the pandemic, suggesting that in these periods, a linear approximation misses important features of the underlying nonlinear model specification. In tranquil times, the KLD is close to zero. Evidently, BART-based predictions are sufficiently close to the ones obtained from using a linear approximation. This finding is not surprising and commonly found in the literature on nonlinear models. Huber, et al. (2020), for instance, find that the performance gains of BART during the pandemic become large, whereas in normal times, gains are more muted. In such situations, nonparametric approaches quickly adapt to these extreme observations.

Zooming in on variable-specific differences reveals that for inflation we find the largest KLD, whereas for the unemployment rate and GDP growth, KLDs are much smaller. Since fullBART is among the set of the best-performing models for inflation forecasts according to the CRPS measure, this suggests that being flexible on the conditional mean and the full variance-covariance matrix seems to pay off.

Finally, there tends to be more evidence of nonlinearity at the longer forecast horizons. This is consistent with our previous evidence that BART-based models forecast particularly well at longer horizons. From a technical perspective, the larger differences for multi-step forecasts are driven by the fact that we iteratively forecast in both cases. Since higher-order forecasts are nonlinear functions of $\tilde{\boldsymbol{A}}$, the approximation error increases with the forecast horizon.

4.6 The role of financial conditions in tail forecasting

In the previous subsections we have shown that BART can be used to produce accurate tail forecasts. With much of the recent literature emphasizing the role of financial conditions in driving negative tail risks to economic activity (see, e.g., Adrian, Boyarchenko, and Giannone (2019) and Delle Monache, De Polis, and Petrella (2020)), we examine the role of financial conditions in the tail risk forecasts of BART-based specifications. In the interest of brevity, we

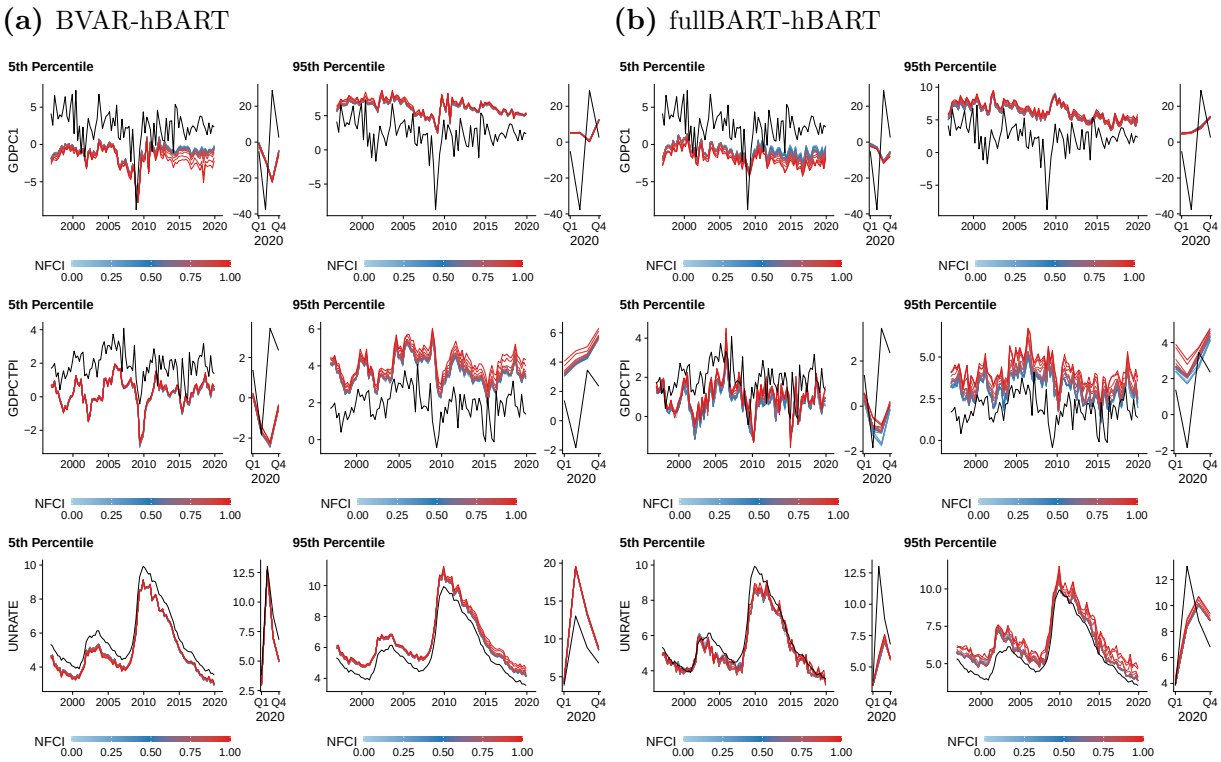

Figure 7: Percentiles of the predictive distributions for different quantiles of the NFCI.
Notes: The quantiles range from 0 to 1 with step size of 0.05. The legend refers to the quantiles.

focus on two models: the BVAR-hBART and fullBART-hBART specifications. This comparison helps to shed light on the role of nonparametric treatments of the conditional variance (hBART) and conditional mean (BART vs. BVAR). For this assessment, we consider NFCI paths over the forecast horizon that are fixed at selected values. These values are the different quantiles of the NFCI, ranging from 0 (the minimum) to 1 (the maximum) with a step-size of 0.05. This provides 21 paths of the NFCI for which we produce conditional forecasts from the models.

4.6.1 Conditional forecasts using fullBART

Figure 7 reports time series of the 5 and 95 percent quantiles of predictive distributions of GDP growth, inflation, and unemployment obtained for each path of the NFCI, over our entire out-of-sample evaluation period (with 2020 separated from the rest of the sample for chart readability). In these charts, blue lines refer to densities conditioning on low values of the NFCI (good financial conditions) and red lines refer to densities conditioning on high values of the NFCI (bad financial conditions). Black lines provide the actual outcomes for growth, inflation, and the unemployment rate.

In the period up to the Great Recession, changes in the NFCI had limited effects on the tails of the predictive density for all variables and both models under consideration. In some

periods (such as the late 1990s) we find that tighter financial conditions have an adverse effect on the left tail of GDP growth. But this finding only holds for fullBART and the magnitudes of the shifts in the conditional distribution are small.

For the period since the Great Recession, conditioning on higher values of the NFCI lowers the 5 percent quantile forecast of GDP growth appreciably. Interestingly, this finding does not carry over to the 95 percent quantile prediction, pointing toward asymmetries in the way financial conditions impact the conditional distribution of output growth. For unemployment, changes in financial conditions have a similar effect: tight financial conditions translate into increases in unemployment, whereas loose financial conditions translate into a tighter labor market and thus lower unemployment rates. It is worth stressing that this effect is more pronounced if we use fullBART as opposed to the BVAR-hBART model.

In the case of inflation, the middle row of the charts indicates that, with both models, conditioning on higher values of the NFCI significantly boosts the 95 percent quantile of the predictive distribution, implying that more adverse conditions are associated with more upside risk to inflation. The NFCI conditioning has relatively little effect on the lower tail forecast for inflation. These patterns for inflation are relatively consistent over the sample, including the pandemic observations.

4.7 Conditional forecasts during the global financial crisis and the pandemic

Focusing on the unemployment rate, Figure 8 reports posterior predictive densities over a few selected quarters from the depths of the Great Recession (2008:Q3 to 2009:Q1) and the 2020 pandemic (2020:Q1 to 2020:Q3). The densities condition on the quantiles of the NFCI that range from 0 (the minimum, blue lines) to 1 (the maximum, red lines).

In broad terms, the charts in the top two rows show that, with the conditional mean taking the linear form of the BVAR, the nonparametric specification of the innovation process through hBART is sufficient to yield predictive distributions that are non-Gaussian. For example, in 2008:Q3, some of the distributions have fat tails, whereas in 2020:Q3, the distributions are sharply peaked rather than bell-shaped. More specifically, in the case of BVAR forecasts during the Great Recession, conditioning on different values of the NFCI impacts the predictive distributions mostly by increasing one of the tails or widening the distributions, with little effect on the mode of the distribution. In 2020, conditioning on different NFCI values has little effect on forecasts for 2020:Q2 and 2020:Q3 but sharply affects the predictive distributions for

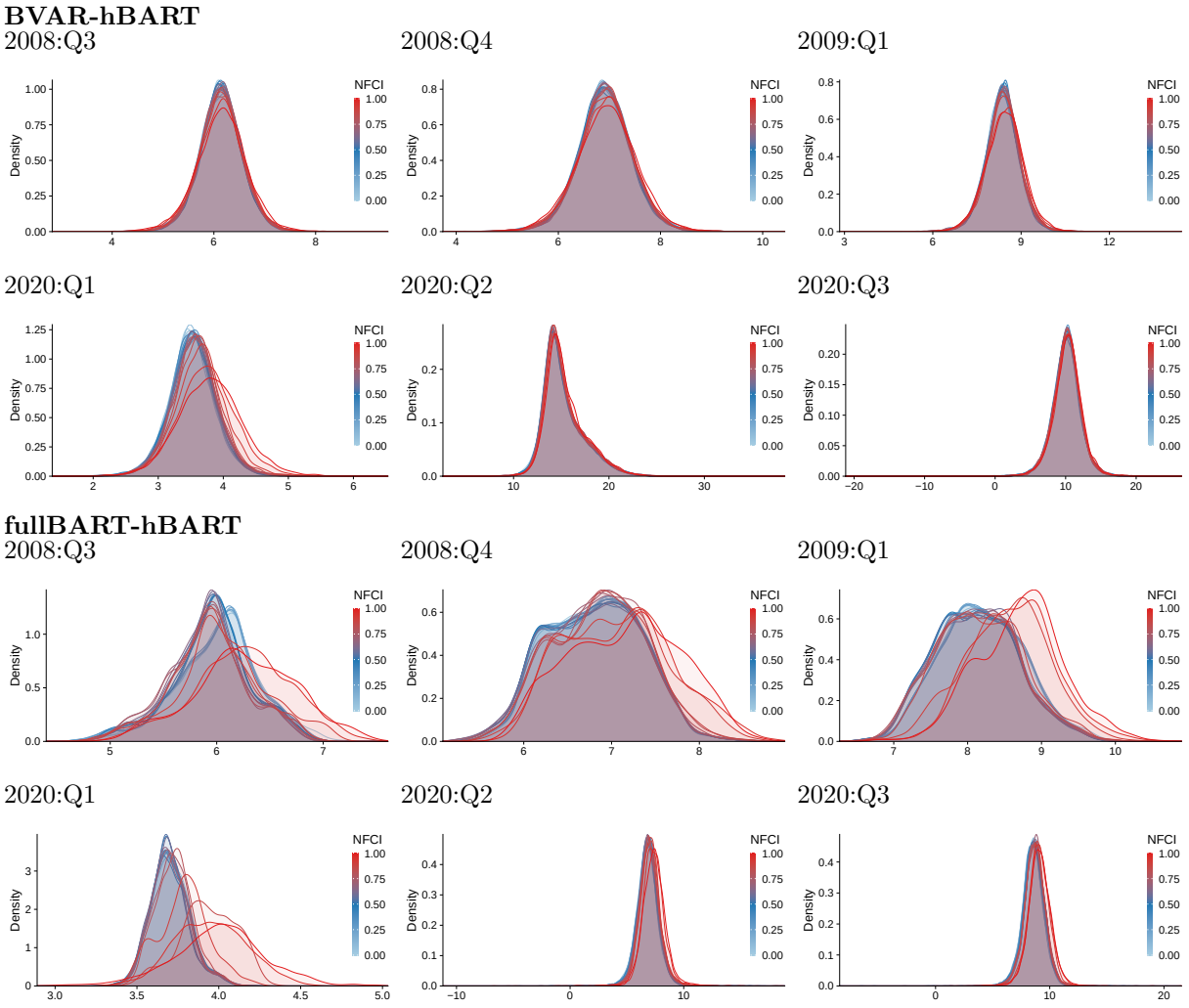

Figure 8: One-step-ahead predictive distributions for unemployment for different values of NFCI.

Notes: The quantiles range from 0 to 1 with step size of 0.05. The legend refers to the quantiles.

2020:Q1, with higher values of the NFCI associated with predictive distributions shifted to the right and widened.

Conditioning on different financial settings has much larger effects on predictive distributions from the fullBART-hBART specification. In this case, pairing a nonparametric specification of the conditional mean with a nonparametric specification of the conditional variance can yield sharply non-Gaussian distributions, with fat tails, asymmetries, or even multi-modality.

Multi-modalities in the predictive distribution for the unemployment rate are most evident from 2008:Q3 through 2009:Q1. In 2008:Q4 and 2009:Q1, even under favorable (near-zero) values of the NFCI, the predictive distributions from the fullBART specification show two clear peaks. As the conditioning NFCI values are increased, the predictive mean shifts some but the variance rises considerably, with much wider distributions. This specification appears to feature these

multi-modalities while yielding favorable forecast accuracy over the full hold-out period.

In the period of the pandemic, the BART-based patterns for 2020:Q2 and 2020:Q3 are similar to the BVAR-based patterns. In these periods, conditioning on different NFCI values has relatively little effect on the predictive distributions. But in 2020:Q1 forecasts, the NFCI has a much greater impact on the predictive distributions, resembling that seen in the Great Recession quarters of 2008:Q3 through 2009:Q1, with higher values of the NFCI sharply raising the predictive mean and variance of the unemployment rate.

5 Concluding remarks

In this paper we have made three main contributions. First, we have used Bayesian additive regression trees (BART) to introduce novel multivariate models that posit nonlinear relationships among macroeconomic variables, their lags, and possibly the lags of the errors. The errors can be either homoskedastic or heteroskedastic, and in the latter case, we consider both a standard stochastic volatility specification and a novel nonparametric specification. The flexible specifications for the conditional mean and variance could be particularly helpful in the presence of parameter time variation and/or for density and tail forecasting.

Second, we have developed MCMC estimation algorithms for each (homoskedastic and heteroskedastic) BART specification. The algorithms are easily scalable to large dimension and thus allow for estimating large semi- and nonparametric VAR models.

Finally, we have evaluated the real-time forecasting performance for a set of US macroeconomic and financial indicators of the various BART models, using a variety of loss functions and a BVAR-SV model as a (strong) benchmark, in addition to a TVP-VAR-SV and a Bayesian quantile regression. The main findings are that when using BART to accommodate nonlinearities, it is less important to allow for heteroskedasticity; the out-of-sample predictive density charts do not show much downside risk asymmetry; and BART specifications can deliver more accurate tail forecasts than BVAR-SV, in particular for inflation.

Overall, the models we develop represent an important addition to the toolbox of empirical macroeconomists and forecasters, due to their flexibility, range of applicability, ease of implementation, and good empirical performance.

In this paper, we have focused on approximating the unknown functions F and G using BART due to its excellent empirical properties. However, there exist several alternative techniques such as Gaussian process and kernel regressions (Quinonero-Candela and Rasmussen

(2005) and Adrian, Boyarchenko, and Giannone (2021)), spline-based models (Shin, Bhattacharya, and Johnson (2020)), or infinite mixtures (Kalli and Griffin (2018)) to flexibly model the conditional mean in a multivariate time series model. Assessing whether these techniques can be used to improve forecasts would be a fruitful avenue of further research, as well as using these more sophisticated models for the identification of structural shocks and their propagation.

References

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

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019), "Vulnerable growth," *American Economic Review*, 109, 1263–89, https://doi.org/10.1257/aer.20161923.

——— (2021), "Multimodality in macrofinancial dynamics," *International Economic Review*, 62, 861–886, https://doi.org/10.1111/iere.12501.

Adrian, Tobias, Federico Grinberg, Nellie Liang, and Sheheryar Malik (2022), "The term structure of growth-at-risk," *American Economic Journal: Macroeconomics*, 14, 283–323, https://doi.org/10.1257/mac.20180428.

Aguilar, Omar, and Mike West (2000), "Bayesian dynamic factor models and portfolio allocation," *Journal of Business & Economic Statistics*, 18, 338–357, https://doi.org/10.1080/07350015.2000.10524875.

Banbura, Marta, Domenico Giannone, and Lucrezia Reichlin (2010), "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92, https://doi.org/10.1002/jae.1137.

Bollerslev, Tim (1990), "Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model," *The Review of Economics and Statistics*, 498–505, https://doi.org/10.2307/2109358.

Breiman, Leo (2001), "Random forests," *Machine Learning*, 45, 5–32, https://doi.org/10.1023/A:1010933404324.

Caldara, Dario, Chiara Scotti, and Molin Zhong (2021), "Macroeconomic and financial risks: A tale of mean and volatility," International Finance Discussion Paper 1326, Board of Governors of the Federal Reserve System, https://doi.org/10.17016/IFDP.2021.1326.

Carriero, Andrea, Joshua Chan, Todd E. Clark, and Massimiliano Marcellino (2022a), "Corrigendum to: Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors," *Journal of Econometrics*, 227, 506–512, https://doi.org/10.1016/j.jeconom.2021.11.010.

Carriero, Andrea, Todd E Clark, and Massimiliano Marcellino (2015), "Bayesian VARs: specification choices and forecast accuracy," *Journal of Applied Econometrics*, 30, 46–73, https://doi.org/10.1002/jae.2315.

Carriero, Andrea, Todd E. Clark, and Massimiliano Marcellino (2016), "Common drifting volatility in large Bayesian VARs," *Journal of Business & Economic Statistics*, 34, 375–390, https://doi.org/10.1080/07350015.2015.1040116.

——— (2019), "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors," *Journal of Econometrics*, 212, 137–154, https://doi.org/10.1016/j.jeconom.2019.04.024.

——— (2020), "Capturing macroeconomic tail risks with Bayesian vector autoregressions," *Federal Reserve Bank of Cleveland Working Paper*, 20-02, https://doi.org/10.26509/frbc-wp-202002.

——— (forthcoming), "Nowcasting tail risk to economic activity at a weekly frequency," *Journal of Applied Econometrics*, https://doi.org/10.1002/jae.2903.

Carriero, Andrea, Todd E. Clark, Massimiliano Marcellino, and Elmar Mertens (2022b), "Addressing COVID-19 outliers in BVARs with stochastic volatility," *The Review of Economics and Statistics*, forthcoming, https://doi.org/10.1162/rest_a_01213.

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

Chan, Joshua CC (2021), "Minnesota-type adaptive hierarchical priors for large Bayesian VARs," *International Journal of Forecasting*, 37, 1212–1226, https://doi.org/10.1016/j.ijforecast.2021.01.002.

Chib, Siddhartha, and Edward Greenberg (1994), "Bayes inference in regression models with ARMA (p, q) errors," *Journal of Econometrics*, 64, 183–206, https://doi.org/10.1016/0304-4076(94)90063-9.

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

——— (2010), "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, 4, 266–298, https://doi.org/10.1214/09-AOAS285.

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

Clark, Todd E, Florian Huber, Gary Koop, Massimiliano Marcellino, and Michael Pfarrhofer (2021), "Investigating growth at risk using a multi-country non-parametric quantile factor model," *arXiv*, 2110.03411.

Clark, Todd E., and Francesco Ravazzolo (2015), "Macroeconomic forecasting performance under alternative specifications of time-varying volatility," *Journal of Applied Econometrics*, 30, 551–575, https://doi.org/10.1002/jae.2379.

Cook, Thomas, and Taeyoung Doh (2019), "Assessing macroeconomic tail risks in a data-rich environment," *Federal Reserve Bank of Kansas City Research Working Paper*, 19-12, https://doi.org/10.18651/RWP2019-12.

Crawford, Lorin, Kris C. Wood, Xiang Zhou, and Sayan Mukherjee (2018), "Bayesian approximate kernel regression with variable selection," *Journal of the American Statistical Association*, 113, 1710–1721, https://doi.org/10.1080/01621459.2017.1361830.

De Nicolò, Gianni, and Marcella Lucchetta (2017), "Forecasting tail risks," *Journal of Applied Econometrics*, 32, 159–170, https://doi.org/10.1002/jae.2509.

Delle Monache, Davide, Andrea De Polis, and Ivan Petrella (2020), "Modeling and forecasting macroeconomic downside risk," *CEPR Discussion Paper Series*, 15109.

Diebold, Francis X., and Robert S. Mariano (1995), "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, 13, 253–263, https://doi.org/10.2307/1392185.

Engle, Robert (2002), "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business & Economic Statistics*, 20, 339–350, https://doi.org/10.1198/073500102288618487.

Ferrara, Laurent, Matteo Mogliani, and Jean-Guillaume Sahuc (2022), "High-frequency monitoring of growth at risk," *International Journal of Forecasting*, 38, 582–595, https://doi.org/10.1016/j.ijforecast.2021.06.010.

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

Galbraith, John W., and Simon van Norden (2019), "Asymmetry in unemployment rate forecast errors," *International Journal of Forecasting*, 35, 1613–1626, https://doi.org/10.1016/j.ijforecast.2018.11.006.

Gefang, Deborah, Gary Koop, and Aubrey Poon (2022), "Forecasting using variational Bayesian inference in large vector autoregressions with hierarchical shrinkage," *International Journal of Forecasting*, in press, https://doi.org/10.1016/j.ijforecast.2021.11.012.

Ghysels, Eric, Leonardo Iania, and Jonas Striaukas (2018), "Quantile-based inflation risk models," *National Bank of Belgium Research Working Paper*, 349.

Giacomini, Raffaella, and Ivana Komunjer (2005), "Evaluation and combination of conditional quantile forecasts," *Journal of Business & Economic Statistics*, 23, 416–431, https://doi.org/10.1198/073500105000000018.

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

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2015), "Prior selection for vector autoregressions," *The Review of Economics and Statistics*, 97, 436–451, https://doi.org/10.1162/rest_a_00483.

Giglio, Stefano, Bryan Kelly, and Seth Pruitt (2016), "Systemic risk and the macroeconomy: An empirical evaluation," *Journal of Financial Economics*, 119, 457–471, https://doi.org/10.1016/j.jfineco.2016.01.010.

Gneiting, Tilmann, and Adrian E. Raftery (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378, https://doi.org/10.1198/016214506000001437.

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

González-Rivera, Gloria, Javier Maldonado, and Esther Ruiz (2019), "Growth in stress," *International Journal of Forecasting*, 35, 948–966, https://doi.org/10.1016/j.ijforecast.2019.04.006.

Goulet Coulombe, Philippe (2020), "The macroeconomy as a random forest," *arXiv*, 2006.12724.

Goulet Coulombe, Philippe, Maxime Leroux, Dalibor Stevanovic, and Stephane Surprenant (2020), "How is machine learning useful for macroeconomic forecasting?" *arXiv*, 2008.12477.

Goulet Coulombe, Philippe, Massimiliano Marcellino, and Dalibor Stevanovic (2021), "Can machine learning catch the COVID-19 recession?" *arXiv*, 2103.01201.

Hernández, Belinda, Adrian E Raftery, Stephen R Pennington, and Andrew C Parnell (2018), "Bayesian additive regression trees using Bayesian model averaging," *Statistics and Computing*, 28, 869–890, https://doi.org/10.1007/s11222-017-9767-1.

Huber, Florian, and Martin Feldkircher (2019), "Adaptive shrinkage in Bayesian vector autoregressive models," *Journal of Business & Economic Statistics*, 37, 27–39, https://doi.org/10.1080/07350015.2016.1256217.

Huber, Florian, Gary Koop, and Luca Onorante (2021), "Inducing sparsity and shrinkage in time-varying parameter models," *Journal of Business & Economic Statistics*, 39, 669–683, https://doi.org/10.1080/07350015.2020.1713796.

Huber, Florian, Gary Koop, Luca Onorante, Michael Pfarrhofer, and Josef Schreiner (2020), "Nowcasting in a pandemic using non-parametric mixed frequency VARs," *Journal of Econometrics*, in press, https://doi.org/10.1016/j.jeconom.2020.11.006.

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

Kalli, Maria, and Jim E. Griffin (2018), "Bayesian nonparametric vector autoregressive models," *Journal of Econometrics*, 203, 267–282, https://doi.org/10.1016/j.jeconom.2017.11.009.

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

Kastner, Gregor, and Florian Huber (2020), "Sparse Bayesian vector autoregressions in huge dimensions," *Journal of Forecasting*, 39, 1142–1165, https://doi.org/10.1002/for.2680.

Kiley, Michael T. (2022), "Unemployment risk," *Journal of Money, Credit and Banking*, forthcoming, https://doi.org/10.1111/jmcb.12888.

Koop, Gary (2013), "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28, 177–203, https://doi.org/10.1002/jae.1270.

Korobilis, Dimitris (2017), "Quantile regression forecasts of inflation under model uncertainty," *International Journal of Forecasting*, 33, 11–20, https://doi.org/10.1016/j.ijforecast.2016.07.005.

Korobilis, Dimitris, and Davide Pettenuzzo (2019), "Adaptive hierarchical priors for high-dimensional vector autoregressions," *Journal of Econometrics*, 212, 241–271, https://doi.org/10.1016/j.jeconom.2019.04.029.

Kozumi, Hideo, and Genya Kobayashi (2011), "Gibbs sampling methods for Bayesian quantile regression," *Journal of Statistical Computation and Simulation*, 81, 1565–1578, https://doi.org/10.1080/00949655.2010.496117.

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

Manzan, Sebastiano (2015), "Forecasting the distribution of economic variables in a data-rich environment," *Journal of Business & Economic Statistics*, 33, 144–164, https://doi.org/10.1080/07350015.2014.937436.

Manzan, Sebastiano, and Dawit Zerom (2013), "Are macroeconomic variables useful for forecasting the distribution of US inflation?" *International Journal of Forecasting*, 29, 469–478, https://doi.org/10.1016/j.ijforecast.2013.01.005.

——— (2015), "Asymmetric quantile persistence and predictability: the case of US inflation," *Oxford Bulletin of Economics and Statistics*, 77, 297–318, https://doi.org/10.1111/obes.12065.

Masini, Ricardo P., Marcelo C. Medeiros, and Eduardo F. Mendes (2021), "Machine learning advances for time series forecasting," *arXiv*, 2012.12802.

Medeiros, Marcelo C., Gabriel F.R. Vasconcelos, Álvaro Veiga, and Eduardo Zilberman (2021), "Forecasting inflation in a data-rich environment: the benefits of machine learning methods," *Journal of Business & Economic Statistics*, 39, 98–119, https://doi.org/10.1080/07350015.2019.1637745.

Mitchell, James, Aubrey Poon, and Gian Luigi Mazzi (2022), "Nowcasting euro area GDP growth using Bayesian quantile regression," in *Essays in Honor of M. Hashem Pesaran: Prediction and Macro Modeling* eds. by Alexander Chudik, Cheng Hsiao, and Allan Timmermann, Advances in Econometrics, vol. 43A: Emerald Publishing Limited, Bingley, 51–72, https://doi.org/10.1108/S0731-90532021000043A004.

Omori, Yasuhiro, Siddhartha Chib, Neil Shephard, and Jouchi Nakajima (2007), "Stochastic volatility with leverage: Fast and efficient likelihood inference," *Journal of Econometrics*, 140, 425–449, https://doi.org/10.1016/j.jeconom.2006.07.008.

Plagborg-Møller, Mikkel, Lucrezia Reichlin, Giovanni Ricco, and Thomas Hasenzagl (2020), "When is growth at risk?" *Brookings Papers on Economic Activity*, Spring, 167–229, https://doi.org/10.1353/eca.2020.0002.

Pratola, Matthew T., Hugh A. Chipman, Edward I. George, and Robert E. McCulloch (2020), "Heteroscedastic BART via multiplicative regression trees," *Journal of Computational and Graphical Statistics*, 29, 405–417, https://doi.org/10.1080/10618600.2019.1677243.

Primiceri, Giorgio E (2005), "Time varying structural vector autoregressions and monetary policy," *The Review of Economic Studies*, 72, 821–852, https://doi.org/10.1111/j.1467-937x.2005.00353.x.

Quinonero-Candela, Joaquin, and Carl Edward Rasmussen (2005), "A unifying view of sparse approximate Gaussian process regression," *The Journal of Machine Learning Research*, 6, 1939–1959.

Reichlin, Lucrezia, Giovanni Ricco, and Thomas Hasenzagl (2020), "Financial variables as predictors of real growth vulnerability," *Deutsche Bundesbank Discussion Paper*, 05/2020.

Ročková, Veronika, and Enakshi Saha (2019), "On theory for BART," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2839–2848, PMLR.

Sentana, Enrique (1995), "Quadratic ARCH models," *The Review of Economic Studies*, 62, 639–661, https://doi.org/10.2307/2298081.

Shin, Minsuk, Anirban Bhattacharya, and Valen E. Johnson (2020), "Functional horseshoe priors for subspace shrinkage," *Journal of the American Statistical Association*, 115, 1784–1797, https://doi.org/10.1080/01621459.2019.1654875.

Stock, James H, and Mark W Watson (2012), "Generalized shrinkage methods for forecasting using many predictors," *Journal of Business & Economic Statistics*, 30, 481–493, https://doi.org/10.1080/07350015.2012.715956.

Wasserman, Larry (2006), *All of Nonparametric Statistics*: Springer Science & Business Media, https://doi.org/10.1007/0-387-30623-4.
West, Kenneth D. (1996), "Asymptotic inference about predictive ability," *Econometrica*, 64, 1067–1084, https://doi.org/10.2307/2171956.
Yu, Keming, and Rana A Moyeed (2001), "Bayesian quantile regression," *Statistics & Probability Letters*, 54, 437–447, https://doi.org/10.1016/S0167-7152(01)00124-9.

A Data Appendix

Table A.1 lists the variables we use (alongside codes and transformations). With one exception, all models used in this paper use all of these variables. The one exception is the small TVP-VAR-SV, which uses only GDP growth, inflation, the federal funds rate, unemployment, and the NFCI. In an early version of this paper, which appeared as Federal Reserve Bank of Cleveland Working Paper 21-08, we considered data sets of different dimension. The reader is referred to that paper for an investigation of our BART-based methods in smaller models. Data are obtained from fred.stlouisfed.org.[17]

Table A.1: Data, description, and information set.

FRED-Code	**Series**	**Trans.**
GDPC1	Real gross domestic product (GDP)	$400\Delta \ln$
GDPCTPI	GDP price index	$400\Delta \ln$
FEDFUNDS	Federal funds rate	level
UNRATE	Unemployment rate	Δ
CPIAUCSL	Consumer price index (CPI)	$400\Delta \ln$
PPIACO	Producer price index (PPI) for all commodities	$400\Delta \ln$
INDPRO	Industrial production	$400\Delta \ln$
PAYEMS	Payroll employment	$400\Delta \ln$
CES0800000001	Payroll employment, services	$400\Delta \ln$
PCECC96	Real personal consumption expenditures	$400\Delta \ln$
A008RA3Q086SBEA	Gross private domestic fixed investment: Nonres.	$400\Delta \ln$
A011RA3Q086SBEA	Gross private domestic fixed investment: Res.	$400\Delta \ln$
PCECTPI	PCE chain price index	$400\Delta \ln$
GPDICTPI	Gross private domestic investment price index	$400\Delta \ln$
CUMFNS	Capacity utilization, manufacturing	level
HOANBS	Nonfarm business sector: Hours of all persons	$400\Delta \ln$
COMPRNFB	Nonfarm bus. sector: Real compensation per hour	$400\Delta \ln$
GS1	1-Year Treasury bond yield	level
GS5	5-Year Treasury bond yield	level
EXUSUK	US / UK exchange rate	$400\Delta \ln$
M2REAL	Real M2 money stock	$400\Delta \ln$
SP500	S&P 500	$400\Delta \ln$
NFCI	Chicago Fed index of financial conditions	level

Notes: "FRED-Code" refers to the code of the respective series at fred.stlouisfed.org. Transformations ("Trans."): Δ indicates first differences and ln is the natural logarithm.

With real-time data vintages available beginning with 1996:Q4, our real-time forecast sample begins with 1997:Q1 and ends with 2020:Q4. If release frequency is higher than quarterly, we use the final vintage per respective quarter for producing forecasts. However, for some variables, real-time data vintages begin later in the sample. In these cases, we use the first vintage to fill in artificial vintages for earlier years, truncating it according to the release calendar. In all cases, the data sample for model estimation starts with 1973:Q2. In evaluating forecasts,

[17] Because the S&P 500 index of stock prices is unavailable prior to 2011 in the online FRED database, we obtained data for this series prior to 2011 from the compiled "FRED-QD" data set, also available from the St. Louis Fed's website.

we measure the actual values of the variables as those of the final available vintage, which is 2021:Q1. Our data set, thus, includes observations during the pandemic. The usefulness of BART for pandemic forecasting was established in previous work by HKOPS. Results using a sample that ends in 2019:Q4 are available in the earlier version of this paper, Federal Reserve Bank of Cleveland Working Paper 21-08. Results differ little from those in the current version of the paper using a data set that includes the pandemic period.

B Technical Appendix

B.1 Priors on the remaining model parameters

On the VAR coefficients $\boldsymbol{A}$ we use a horseshoe prior (Carvalho, Polson, and Scott (2010)). This choice is motivated by two main reasons. First, global-local shrinkage priors such as the horseshoe possess excellent forecasting properties in high dimensions (see, e.g., Huber and Feldkircher, 2019; Korobilis and Pettenuzzo, 2019; Chan, 2021). Second, and this is crucial for our extensive real-time forecasting exercise, the horseshoe prior does not rely on a single hyperparameter. This implies that no cross-validation is necessary.

Let $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iK})'$ denote the i^{th} row of $\boldsymbol{A}$ and a_{ij} the i^{th} element of $\boldsymbol{a}_i$. The horseshoe prior is a hierarchical Gaussian prior on a_{ij}:

$$a_{ij}|\lambda_i, \psi_{ij} \sim \mathcal{N}(0, \psi_{ij}^2\lambda_i^2), \quad \psi_{ij} \sim \mathcal{C}^+(0,1), \quad \lambda_i \sim \mathcal{C}^+(0,1), \tag{B.1}$$

with λ_i being a shrinkage hyperparameter that applies to all coefficients in equation i and ψ_{ij} denotes a coefficient-specific scaling parameter. Both λ_i and ψ_{ij} feature a half-Cauchy prior $\mathcal{C}^+$. This prior belongs to the general class of global-local shrinkage priors that shrink globally (through λ_i) but allow for local deviations (through ψ_{ij}) if λ_i is close to zero. For the factor loadings in $\boldsymbol{\Lambda}$ we use a horseshoe prior similar to the one in (B.1). The only exception is that the global shrinkage parameter applies to each column of $\boldsymbol{\Lambda}$. This allows for pushing coefficients associated with irrelevant factors to zero.

In the case where we use a model that features stochastic volatility, the prior on the unconditional mean is Gaussian with mean zero and variance 10; the prior on the persistence parameter, denoted by ρ_i, is Beta distributed $\frac{\rho_i+1}{2} \sim \mathcal{B}(25,5)$; and the prior on the variance of the log-volatility process is Gamma distributed $\mathcal{G}(1/2, 1/2)$.

If we use a model with homoskedastic shocks we use an inverse Gamma prior on the main

diagonal elements of $\boldsymbol{H}_t$, σ_i^2, which we set to be rather uninformative, i.e., $\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, c_1)$. The hyperparameters c_0, c_1 are set equal to 0.01.

B.1.1 Sampling the remaining unknowns of the model

Conditional on the trees and the error volatilities, one can sample the VAR coefficients and the covariance parameters in a single block using standard textbook results for the linear regression model. The full conditional posterior of $\boldsymbol{a}_i$ is multivariate Gaussian:

$$\begin{aligned}
\boldsymbol{a}_i|\bullet &\sim \mathcal{N}(\overline{\boldsymbol{\beta}}_i, \overline{\boldsymbol{V}}_i), \\
\overline{\boldsymbol{V}}_i &= (\tilde{\boldsymbol{X}}_i'\tilde{\boldsymbol{X}}_i + \underline{\boldsymbol{V}}^{-1})^{-1}, \\
\overline{\boldsymbol{\beta}}_i &= \overline{\boldsymbol{V}}_i\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{y}}_i,
\end{aligned}$$

where $\tilde{\boldsymbol{X}}_i'$ denotes a $T \times K$ matrix with typical t^{th} row $\tilde{\boldsymbol{x}}_{it}' = \boldsymbol{x}_t'/e^{h_{it}/2}$, $\tilde{\boldsymbol{y}}_i$ has typical t^{th} element $(y_{it} - g_i(\boldsymbol{z}_t) - \boldsymbol{\lambda}_i\boldsymbol{\delta}_t)/e^{h_{it}/2}$, and $\underline{\boldsymbol{V}}$ denotes a diagonal prior variance-covariance matrix constructed using the variances described in (B.1). The $\bullet$ notation indicates that we condition on the remaining parameters of the model.

Notice that if we use a model that assumes f_i and g_i to be nonlinear, this sampling step can be omitted. In the case where we estimate the errorBART model, $\boldsymbol{x}_{it}$ will be replaced with $(\boldsymbol{\eta}_{t-1}', \ldots, \boldsymbol{\eta}_{t-p}')'$.

The factor loadings $\boldsymbol{\Lambda}$ are simulated from a full conditional posterior that is conditionally Gaussian and takes a standard form. For each row of $\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}_i$, we simulate from:

$$\begin{aligned}
\boldsymbol{\lambda}_i'|\bullet &\sim \mathcal{N}(\overline{\boldsymbol{\lambda}}_i, \overline{\boldsymbol{W}}_i), \\
\overline{\boldsymbol{W}}_i &= (\tilde{\boldsymbol{F}}_i'\tilde{\boldsymbol{F}}_i + \underline{\boldsymbol{W}}_i^{-1})^{-1}, \\
\overline{\boldsymbol{\lambda}}_i &= \overline{\boldsymbol{W}}_i(\tilde{\boldsymbol{F}}_i'\hat{\boldsymbol{y}}_i).
\end{aligned}$$

$\tilde{\boldsymbol{F}}_i$ has a typical t^{th} row $\boldsymbol{\delta}_t/e^{h_{it}/2}$ and the t^{th} element of $\hat{\boldsymbol{y}}_i$ is $(y_{it} - f_i(\boldsymbol{x}_t) - g_i(\boldsymbol{z}_t))/e^{h_{it}/2}$ and $\underline{\boldsymbol{W}}_i$ is an $(MQ) \times (MQ)$ prior variance-covariance matrix that we construct using similar quantities to the ones described in (B.1).

The factors are simulated on a t-by-t basis from Gaussian distributions. The full conditionals are given by (see Aguilar and West (2000)):

$$\boldsymbol{\delta}_t|\bullet \sim \mathcal{N}(\boldsymbol{\zeta}_t(\boldsymbol{y}_t - f(\boldsymbol{x}_t) - g(\boldsymbol{z}_t)), \boldsymbol{\Omega}_t - \boldsymbol{\zeta}_t\boldsymbol{\Delta}\boldsymbol{\zeta}_t'),$$

with $\boldsymbol{\zeta}_t = \boldsymbol{\Omega}_t \boldsymbol{\Lambda}' \boldsymbol{\Delta}$ and $\boldsymbol{\Delta} = \boldsymbol{\Lambda} \boldsymbol{\Omega}_t \boldsymbol{\Lambda}' + \boldsymbol{H}_t$. For each point in time, draws of $\boldsymbol{\delta}_t$ are obtained in a (conditionally) independent manner from this Q-dimensional Gaussian distribution.

In models that include stochastic volatility, we use the efficient sampler outlined in Kastner and Frühwirth-Schnatter (2014). This sampler also exploits the 10-component mixture approximation to the $\log \chi_1^2$ distribution but restates the conditionally Gaussian and linear state space model in terms of a big regression model with the regression coefficients being the log-volatilities. This gives rise to an algorithm that samples the volatilities all without a loop from a $(T-1)$-dimensional multivariate Gaussian distribution.

If we use a homoskedastic model, the error variances can easily be sampled from an inverse Gamma posterior with

$$\sigma_i^2 | \bullet \sim \mathcal{G}^{-1}\left(c_0 + \frac{T}{2}, c_1 + \frac{\sum_{t=1}^T \varepsilon_{it}^2}{2}\right).$$

Finally, the hyperparameters of the horseshoe prior are simulated using the auxiliary sampler proposed in Makalic and Schmidt (2015). We will outline the relevant full conditionals for the prior on $\boldsymbol{a}_i$ only. The hyperparameters for the prior on $\boldsymbol{\Lambda}$ take precisely the same form.

Makalic and Schmidt (2015) introduce auxiliary random variables ζ_i and κ_{ij}. Conditional on these, the posteriors of ψ_i^2 and λ_i^2 are inverse Gamma distributed:

$$\psi_{ij}^2 | \bullet \sim \mathcal{G}^{-1}\left(1, \frac{1}{\kappa_{ij}} + \frac{a_{ij}^2}{2\lambda_i^2}\right), \quad \lambda_i^2 | \bullet \sim \mathcal{G}^{-1}\left(\frac{K+1}{2}, \frac{1}{\zeta_i} + \frac{1}{2}\sum_{j=1}^K \frac{a_{ij}^2}{\psi_{ij}^2}\right),$$

as is the posterior of the auxiliary parameters:

$$\kappa_{ij} | \bullet \sim \mathcal{G}^{-1}\left(1, 1 + \frac{1}{\zeta_i}\right), \quad \zeta_i | \bullet \sim \mathcal{G}^{-1}\left(1, 1 + \frac{1}{\lambda_i^2}\right).$$

B.2 Further competing models

B.2.1 Time-varying-parameter VAR with factor stochastic volatility

Borrowing notation from Eq. (1), the TVP-VAR-SV model is given by:

$$\boldsymbol{y}_t = \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{\Sigma}_t),$$

with $F(\boldsymbol{x}_t) = \boldsymbol{A}_t \boldsymbol{x}_t$ and G omitted. $\boldsymbol{A}_t$ is an $M \times K$-matrix of time-varying coefficients, and $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda} \boldsymbol{\Omega}_t \boldsymbol{\Lambda}' + \boldsymbol{H}_t$ is decomposed using an FSV model as for our other specifications. To

establish the TVPs and our prior setup, define $\tilde{\boldsymbol{y}}_t = \boldsymbol{y}_t - \boldsymbol{\Lambda}\boldsymbol{\delta}_t$, such that

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{A}_t\boldsymbol{x}_t + \boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{H}_t),$$

which allows us to consider the TVP-VAR-SV as independent TVP regressions conditional on $\boldsymbol{\Lambda}\boldsymbol{\delta}_t$. Let $\boldsymbol{a}_{it}$ denote the i^{th} row of $\boldsymbol{A}_t$, and consider equation i:

$$\begin{aligned}\tilde{y}_{it} &= \boldsymbol{x}_t'\boldsymbol{a}_{it} + e_{it}, \quad e_{it} \sim \mathcal{N}(0, e^{h_{it}}),\\ \boldsymbol{a}_{it} &= \boldsymbol{a}_{it-1} + \boldsymbol{\eta}_{it}, \quad \boldsymbol{\eta}_{it} \sim \mathcal{N}(\boldsymbol{0}_K, \boldsymbol{\Upsilon}_i),\end{aligned}$$

with $\boldsymbol{\Upsilon}_i = \text{diag}(\upsilon_{i1}, \ldots, \upsilon_{iK})$ and $\sqrt{\boldsymbol{\Upsilon}_i} = \text{diag}\left(\sqrt{\upsilon_{i1}}, \ldots, \sqrt{\upsilon_{iK}}\right)$. Algorithmically, our implementation relies on the non-centered parameterization of Frühwirth-Schnatter and Wagner (2010):

$$\begin{aligned}\tilde{y}_{it} &= \boldsymbol{a}_{i0}'\boldsymbol{x}_t + \tilde{\boldsymbol{a}}_{it}'\sqrt{\boldsymbol{\Upsilon}_i}\boldsymbol{x}_t + e_{it},\\ \tilde{\boldsymbol{a}}_{it} &= \tilde{\boldsymbol{a}}_{it-1} + \boldsymbol{\eta}_{it}, \quad \boldsymbol{\eta}_{it} \sim \mathcal{N}(\boldsymbol{0}_K, \boldsymbol{I}_K), \quad \tilde{\boldsymbol{a}}_{i0} = \boldsymbol{0}_K.\end{aligned}$$

This splits the TVPs into a random walk process for $\tilde{\boldsymbol{a}}_{it}$ with standard normal errors and a time-invariant part $\boldsymbol{a}_{i0} = (a_{i1,0}, \ldots, a_{iK,0})'$. The square roots of the state innovation variances, $\sqrt{\boldsymbol{\Upsilon}_i}$, are featured in the measurement equation and can be treated as regression coefficients. We specify equation-specific horseshoe priors on both:

$$\begin{aligned}a_{ij,0} &\sim \mathcal{N}(0, \psi_{A,ij}^2\lambda_{A,i}^2), \quad \psi_{A,ij} \sim \mathcal{C}^+(0,1), \quad \lambda_{A,i} \sim \mathcal{C}^+(0,1),\\ \sqrt{\upsilon_{ij}} &\sim \mathcal{N}(0, \phi_{\Upsilon,ij}^2, \lambda_{\Upsilon,i}^2), \quad \psi_{\Upsilon,ij} \sim \mathcal{C}^+(0,1), \quad \lambda_{\Upsilon,i} \sim \mathcal{C}^+(0,1).\end{aligned}$$

This specification is similar to Huber, Koop, and Onorante (2021). We sample the TVPs equation-by-equation using a forward-filtering backward-sampling algorithm, and conditional on these draws, update the other parameters of the model using the horseshoe posteriors provided above. The FSV-part of the model is identical to the one for the BART-based variants with conventional SV to enable direct comparisons.

B.2.2 Bayesian quantile regression

The BQR is based on estimating univariate quantile-specific models with $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_{t-p}')'$ of size K for the focus variables. Each of the quantile regressions also features an intercept,

which we ignore in what follows for brevity.

To simplify notation, we refer to individual variables by y_t and omit the i subscript. For quantile $\tau \in (0,1)$ we have:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_\tau + \epsilon_t, \quad \epsilon_t \sim \text{AL}_\tau(\sigma_\tau). \tag{B.2}$$

The asymmetric Laplace (AL) distribution is chosen as the likelihood due to the arguments provided in Yu and Moyeed (2001). We follow Kozumi and Kobayashi (2011) and rely on an auxiliary representation of the AL distribution:

$$\epsilon_t = \theta_p z_{\tau t} + \tau_\tau \sqrt{\sigma_\tau z_{\tau t}} u_t, \quad u_t \sim \mathcal{N}(0,1), \quad \theta_\tau = \frac{1-2\tau}{\tau(1-\tau)}, \quad \pi_\tau^2 = \frac{2}{\tau(1-\tau)},$$

with $z_{\tau t} \sim \mathcal{E}(\sigma_p)$ following an exponential distribution. Notice that we may write Eq. (B.2) as a conditionally Gaussian model:

$$\tilde{y}_{\tau t} = \tilde{\boldsymbol{x}}_{\tau t}'\boldsymbol{\beta}_\tau + u_t,$$

with $\tilde{y}_{\tau t} = (y_t - \theta_\tau z_{\tau t})/(\pi_\tau \sqrt{\sigma_\tau z_{\tau t}})$ and $\tilde{\boldsymbol{x}}_{p\tau t} = (\tau_p \sqrt{\sigma_p z_{\tau t}} \boldsymbol{I}_K)^{-1} \boldsymbol{x}_t$. Based on conditional Gaussianity, we may use any prior on the quantile-specific coefficients $\boldsymbol{\beta}_\tau$ that one would use in a conventional linear Bayesian regression to design a Gibbs sampling algorithm. In line with our arguments with respect to the other models, we opt for a horseshoe prior on the j^{th} coefficient, $\beta_{j\tau}$, for $j = 1, \ldots, K$:

$$\beta_{j\tau} \sim \mathcal{N}(\psi_{\beta,j\tau}^2 \lambda_{\beta,\tau}^2), \quad \psi_{\beta,j\tau} \sim \mathcal{C}^+(0,1), \quad \lambda_{\beta,\tau} \sim \mathcal{C}^+(0,1),$$

which provides shrinkage by quantile. A similar prior structure for quantile regression has been proposed in Clark, et al. (2021). For the scale parameter of the AL distribution, we use a weakly informative inverse Gamma prior: $\sigma_\tau \sim \mathcal{G}^{-1}(3, 0.3)$.

The corresponding posterior distributions can be found in Kozumi and Kobayashi (2011), and those for the horseshoe prior are shown above in this appendix. For multi-step-ahead forecasts, we specify Eq. (B.2) as a predictive equation with y_{t+h} as the dependent variable.

C Additional Empirical Results for US Data

This appendix includes additional empirical results. The first subsection shows volatility forecasts while the second and third provide additional full sample results such as heatmaps measuring variable relevance and MCMC convergence diagnostics.

C.1 Volatility forecasts

To shed light on the different ways we have of modeling the volatility process, Figure C.1 presents, for each of the three variables, 1- and 4-step-ahead forecasts of the volatility using different models of volatility applied to the BVAR and fullBART specifications of the conditional mean. For both models, we plot lines produced for the three volatility treatments to offer an easy comparison. The first point worth noting is that there are differences in the volatility estimates produced by the three treatments of the error variances. As expected, homoskedastic modeling of error variances tends to produce volatility forecasts that are relatively smoother and at a higher level. These very gradual changes are produced by our recursive forecasting design that implies almost no discounting of past information.

hBART tends to produce volatility forecasts that are similarly smooth, but much lower than the homoskedastic ones. But volatility forecasts by SV models tend to be more volatile than the other approaches. One exception to this pattern is revealing: For the linear BVAR, SV and hBART produce volatility estimates that are quite similar. Informally speaking, in models where both conditional mean and variance are modeled using BART approaches, the model can "choose" to put nonlinearities in the conditional mean or the conditional variance, and the choice made is typically to put them in the conditional mean. In the linear VAR such a choice is not possible and, thus, the hBART estimates put the nonlinearities in the conditional variances in a similar manner to SV. The fact that the homoskedastic version of BART tends to forecast well also supports the idea that most of the benefits of the nonparametric approach are obtained in its modeling of the conditional mean as opposed to the conditional variance. Yet, it is worth mentioning that with BART models we tend to find more evidence of changes in volatility for unemployment than for GDP growth. The latter increases more during and soon after recessions.

There are also interesting differences in the volatility forecasts during the pandemic. Particularly for $h = 1$ and for unemployment and GDP growth, SV models are forecasting much larger increases in volatility than the other approaches. This is consistent with findings in

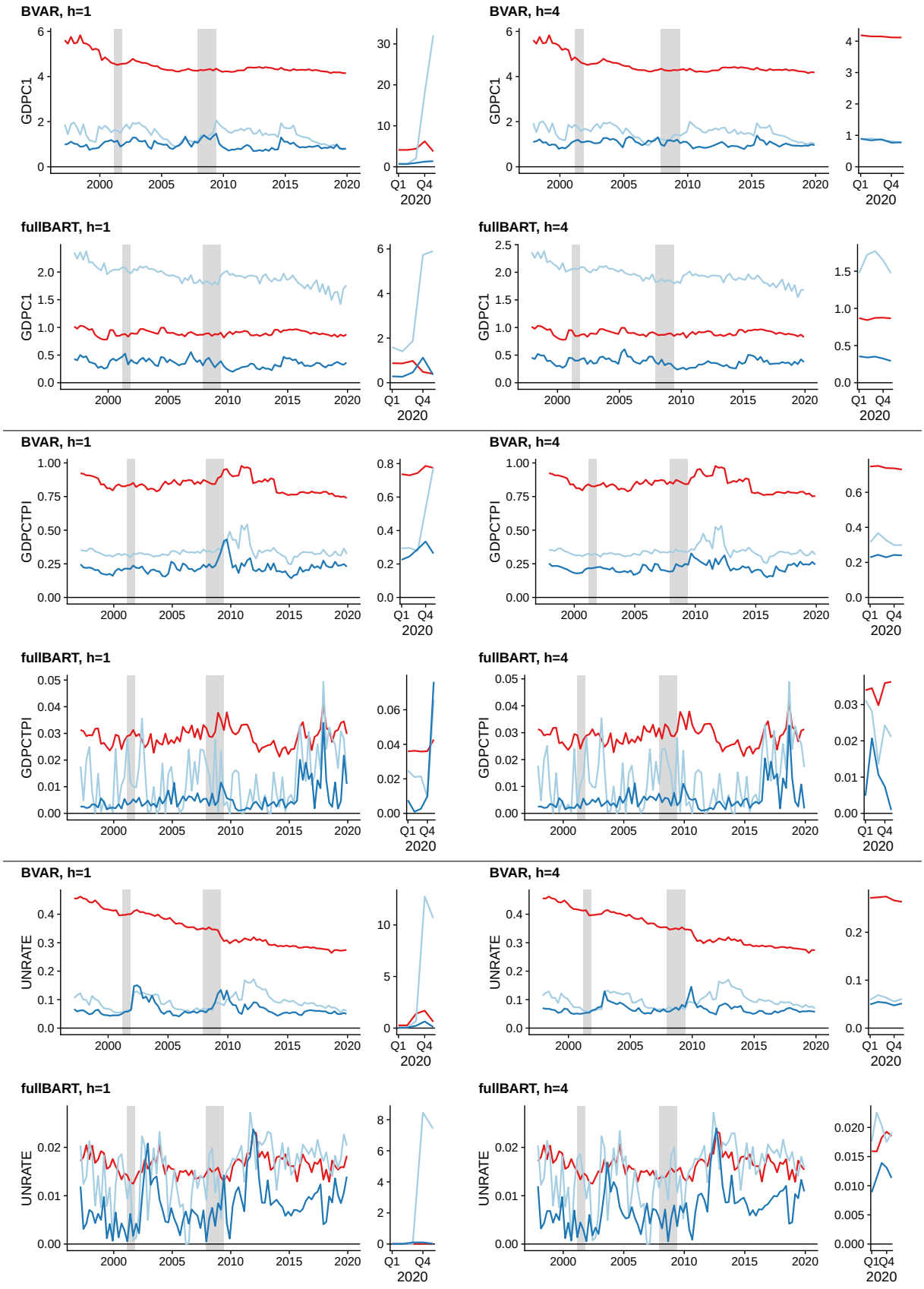

Figure C.1: Volatility predictions (posterior median) for BVAR and fullBART.

Notes: Constant volatility (—), SV (—), hBART (—).

HKOPS, where the extreme pandemic observations had a great impact on the conditional mean that was successfully picked up by BART approaches, leaving less variation in the conditional variance. This pattern is also relevant when the volatility of 1-step-ahead forecast errors is used as a proxy for uncertainty, as the latter would have increased much less with BART than with standard BVAR-SV models.

C.2 In-sample results

In the main text we discuss splitting rules across the three focus variables and for fullBART-hBART only. In Figure C.2 we show similar plots for BART, mixBART and fullBART and all three specifications of the error variances. We drop errorBART from the comparison since this specification includes the lagged reduced-form shocks and a comparison of the remaining three specifications is difficult.

Figure C.2: Percentages of variables in splitting rules relative to the total number of splits.

In principle, the heatmaps tell a story very similar to the one provided in the main text. Irrespective of the variance specification chosen, the percentages of variables showing up in a given splitting rule (normalized by the total number of splits) look similar. This indicates that the nonparametric conditional mean model efficiently extracts information and becoming more flexible on how the error variances evolve over time has only a minor impact on variable relevance.

In the main text we state that the heatmaps draw a picture of which variables drive the

Table C.1: Summary statistics for trees in multivariate BART variants.

	Mean No. of terminal nodes			Max. No. of terminal nodes			Max. obs. in terminal node		
Model	GDPC1	GDPCTPI	UNRATE	GDPC1	GDPCTPI	UNRATE	GDPC1	GDPCTPI	UNRATE
Homoskedastic									
BART	2.24	2.26	2.21	4.81	4.91	4.85	164.77	163.44	169.57
mixBART	2.25	2.26	2.22	4.86	4.92	4.86	164.63	163.33	169.19
errorBART	2.06	2.06	2.06	4.17	4.21	4.18	168.65	168.73	168.89
fullBART	2.25	2.26	2.16	4.91	4.81	4.60	164.77	163.75	172.17
SV									
BART	2.22	2.26	2.18	4.86	4.86	4.68	166.13	163.75	169.72
mixBART	2.22	2.26	2.18	4.82	4.93	4.73	165.71	163.59	169.54
errorBART	2.07	2.07	2.07	4.18	4.18	4.23	170.49	170.16	171.84
fullBART	2.23	2.27	2.17	4.82	4.90	4.58	164.82	162.60	172.12
hBART									
BART	2.23	2.27	2.19	4.83	4.94	4.79	166.04	163.22	169.29
mixBART	2.22	2.27	2.19	4.81	4.99	4.79	165.80	163.41	169.53
errorBART	2.06	2.06	2.06	4.16	4.17	4.19	169.03	168.91	169.19
fullBART	2.21	2.25	2.14	4.71	4.86	4.47	166.34	163.82	171.98

Notes: "Mean No. of terminal nodes" refers to the number of terminal nodes averaged over all S trees and MCMC samples, while "Max. No. of terminal nodes" indicates the maximum number of terminal nodes across all S trees averaged over MCMC draws. "Max. obs. in terminal node" refers to the number of observations (fitted values) allocated to the terminal node that contains the larger number of observations.

conditional mean but do not provide information on how the trees look. Since the number of trees in the BART models is vast, we focus on summary statistics for the sum of trees model. These are reported in Table C.1. The column "Mean No. of terminal nodes" refers to the posterior mean number of terminal nodes averaged over all S trees, while the column "Max. No. of terminal nodes" indicates the posterior mean maximum number of terminal nodes across all S trees.The column "Max. obs. in terminal node" refers to the number of observations (fitted values) allocated to the terminal node that contains the larger number of observations across all trees and iterations of the algorithm. The table shows that trees appear to be rather simple for most models, with a mean (max) number of terminal nodes across equations just over 2 (just below 5). Consistent with most of the findings discussed in the main text (and also for the table showing variable relevance measures) we find no discernible differences in tree complexity across volatility specifications. Again, the full sample results tell a story that once we introduce BART into the the conditional mean, the corresponding specification on the error covariance plays a smaller role and does not substantially impact the conditional mean model.

C.3 MCMC convergence diagnostics

In this section we provide some evidence that our MCMC algorithm is mixing well. This is achieved by considering inefficiency factors that take into account the autocorrelation between successive draws from the joint posterior. Values of the inefficiency factors below 30 are typically viewed to indicate well-mixing chains (see, e.g., Primiceri (2005)). Since our models feature many latent states and parameters, we report averages (over time) of inefficiency factors of draws from

Table C.2: In-sample inefficiency factors for fitted values

	GDPC1			GDPCTPI			UNRATE		
Model/Size	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
BART homosk.	8.8	10.5	6.9	6.6	5.3	4.5	16.8	11.1	7.9
mixBART homosk.	8.7	10.2	7.3	6.6	5.5	4.4	18.1	10.8	7.8
errorBART homosk.	2.3	4.5	7.9	2.6	2.7	10.7	2.6	4.0	11.4
fullBART homosk.	9.2	13.7	11.0	7.3	7.2	11.7	17.0	14.9	12.8
BART SV	20.6	14.0	8.2	8.4	6.5	4.2	23.7	12.1	7.6
mixBART SV	18.8	14.3	8.3	8.0	6.4	4.3	22.9	32.0	8.1
errorBART SV	2.6	4.0	3.0	4.3	4.3	16.4	4.1	3.4	2.8
fullBART SV	2.4	2.3	2.4	10.0	8.7	30.0	22.2	23.0	18.5
BART hBART	21.1	16.0	9.0	9.3	7.4	5.7	26.8	16.8	9.2
mixBART hBART	20.8	14.8	9.2	8.9	7.2	5.8	29.1	17.8	8.7
errorBART hBART	2.6	2.4	12.2	2.8	3.1	13.2	3.0	2.5	26.8
fullBART hBART	36.5	27.2	21.8	10.8	10.8	21.3	29.1	39.4	27.0

the posterior of the conditional mean functions.

These are depicted in Table C.2. The table points toward favorable convergence properties, with draws from the posterior of the latent functions displaying very little autocorrelation. For some equations and models, inefficiency factors are above 2, whereas in the worst cases we obtain inefficiency factors between 30 and 39.4. These numbers signal strong convergence properties, in line with the findings of Chipman, George, and McCulloch (2010), who also provide evidence that the BART components of the MCMC sampler mix well.

D Monte Carlo Evidence

In this appendix, we present the results of a small Monte Carlo study where the data-generating processes (DGPs) are constant parameter and time-varying parameter (TVP) VARs. Both DGPs have SV. The goal is to see how well our various BART models can approximate these parametric models. The DGPs take the form:

$$\boldsymbol{y}_t = \sum_{p=1}^{P} \boldsymbol{A}_{pt} \boldsymbol{y}_{t-p} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t),$$

where $\boldsymbol{A}_{pt}$ are $M \times M$-matrices for lag p with characteristic elements $a_{ij,pt}$. Moreover, we set $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}\boldsymbol{\Omega}_t\boldsymbol{\Lambda} + \boldsymbol{H}_t$, where $\boldsymbol{\Lambda}$ is $M \times Q$, $\boldsymbol{\Omega}_t = \text{diag}(e^{\omega_{1t}}, \dots, e^{\omega_{Qt}})$ and $\boldsymbol{H}_t = \text{diag}(e^{h_{1t}}, \dots, e^{h_{Mt}})$. We consider small and large data sets, with $M \in \{5, 25\}$, with $P = 5$ lags and $Q = 2$ factors, and simulate $T_{\text{full}} = T_{\text{burn}} + T + P$ observations. Subsequently, we discard the initial $T_{\text{burn}} = 50$ observations to mute the effects of the initial conditions and construct the design matrices from $T + P$ observations such that all VARs feature $T = 200$ observations. The corresponding parameters are simulated as follows:

- We simulate the initial conditions of the first-order autoregressive coefficients from $a_{ii,10} \sim \mathcal{N}(0.3, \varsigma_A^2)$; for $i \neq j$, we have $a_{ij,10} \sim \mathcal{N}(0, \varsigma_A^2)$. Coefficients associated with higher-order lags are simulated as $a_{ij,p0} \sim \mathcal{N}(0, \varsigma_A^2/p^2)$, which implies that coefficient matrices become more sparse for distant lags. We set $\varsigma_A^2 = 0.1^2$. Moreover, we assume random walks for $a_{ij,pt}$:

  $$a_{ij,pt} = a_{ij,pt-1} + \vartheta_{ij,p}\eta_t,$$

  and simulate $\vartheta_{ij,p} \sim \mathcal{G}^{-1}(12, 0.03/p)$, again with the amount of time variation decreasing with the lag order p. For the constant parameter DGPs, $\vartheta_{ij,p} = 0$ for all i, j and p. In this case, the VAR coefficients are given by the initial conditions in $\boldsymbol{A}_{p0}$. We only consider stable simulations in terms of the VAR coefficients and re-draw them t-by-t in case they yield explosive multivariate systems.

- The elements of the loadings matrix $\boldsymbol{\Lambda}$, λ_{ij}, are simulated as $\lambda_{ij} \sim \mathcal{N}(0, 0.2^2)$. We consider autoregressive laws of motion for the logarithm of the diagonal elements of $\boldsymbol{\Omega}_t$ and $\boldsymbol{H}_t$:

$$\begin{aligned} \omega_{qt} &= \phi_\omega \omega_{qt-1} + \varsigma_\omega \varepsilon_{qt}, && \text{for } q = 1, \dots, Q \\ h_{it} &= \mu_h + \phi_h(h_{it-1} - \mu_h) + \varsigma_h \varepsilon_{it}, && \text{for } i = 1, \dots, M, \end{aligned}$$

and set $\mu_h = -1$, $\phi_\omega = \phi_h = 0.95$ and $\varsigma_h = \varsigma_\omega = 0.1$.

Figures D.1 and D.2 plot credible intervals for the conditional mean functions in the five equations for five different artificial data sets for fifteen different models (i.e. five specifications for the conditional mean and three for the conditional variance). It can be seen that the various BART models do well in fitting the conditional mean. Of the four different BART approaches for the conditional mean we find errorBART to perform the worst, a finding that also occurs in our empirical results using US data. The homoskedastic linear VAR performs particularly poorly when faced with our DGPs that involve SV.

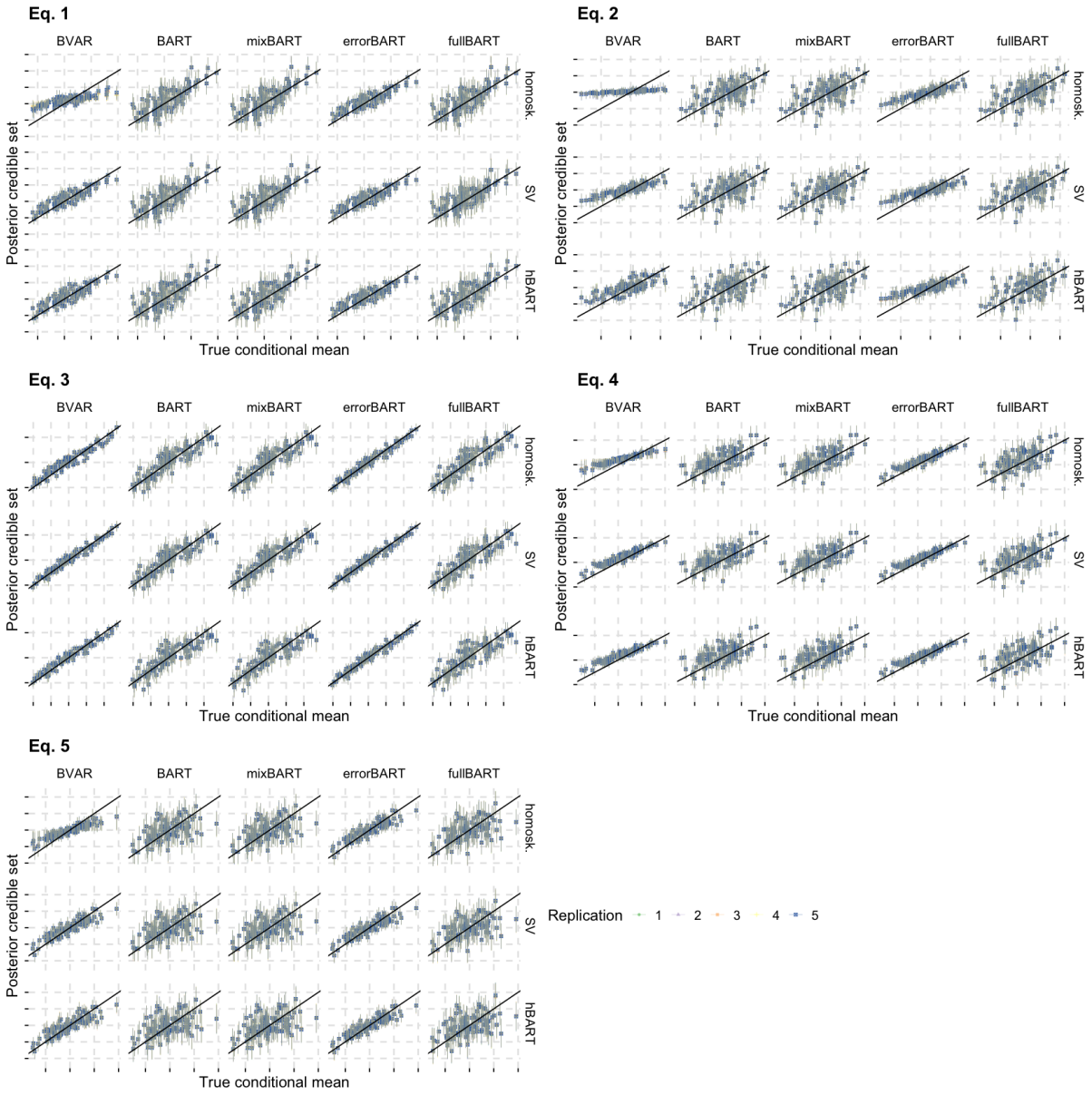

Figure D.1: Posterior credible sets and true conditional mean function for model specifications across five runs of the models for the small constant parameter DGP.

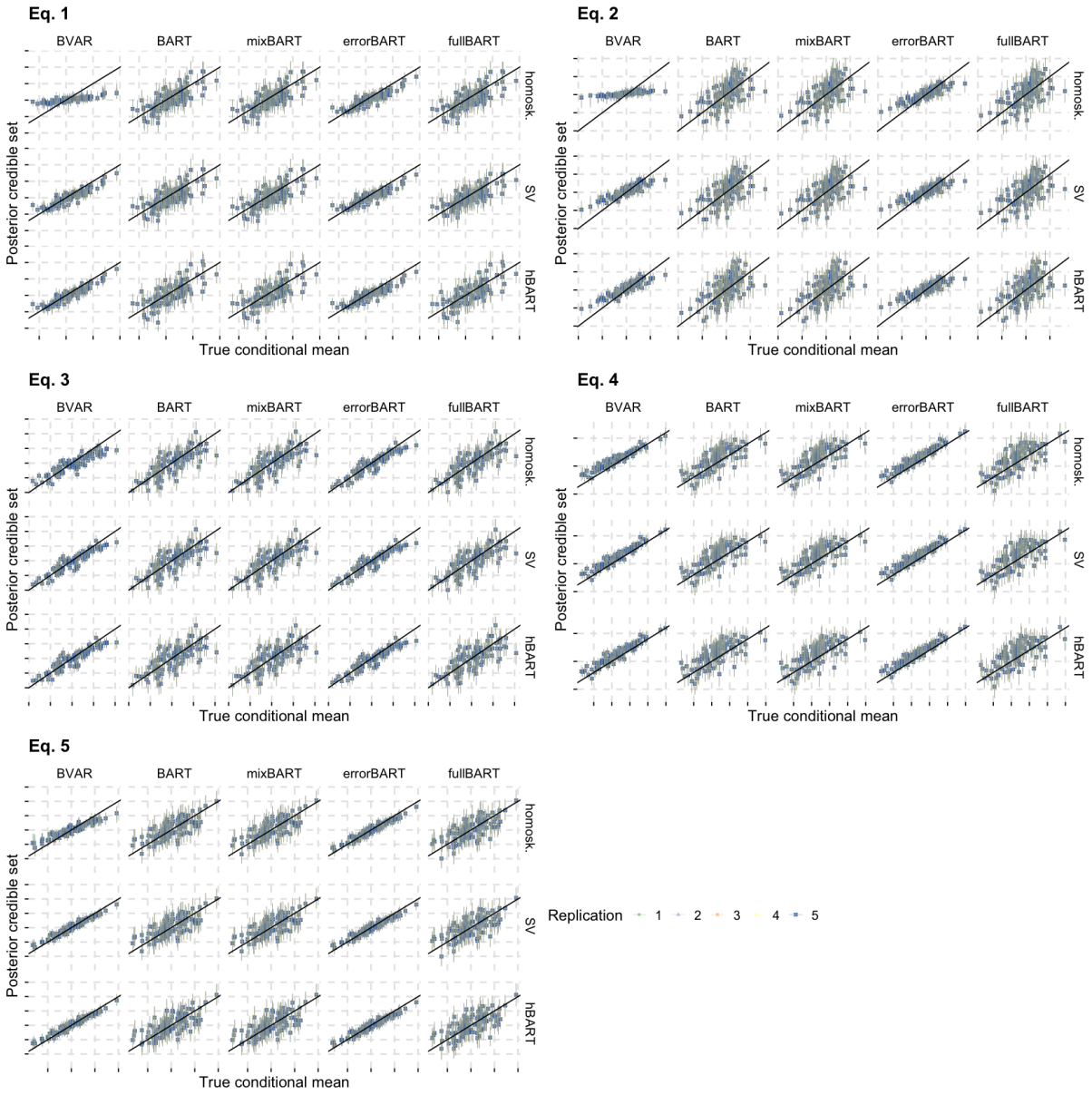

Figure D.2: Posterior credible sets and true conditional mean function for model specifications across five runs of the models for the small TVP DGP.

Tables D.1, D.2, D.3 and D.4 present tree diagnostics, averaged over 50 artificial data sets, for our four DGPs (i.e., two different numbers of variables times two different specifications involving constant and time-varying parameters). These confirm our findings using macroeconomic data that BART tends to use relatively simple tree structures.

Table D.1: Tree diagnostics for the small constant parameter DGPs.

	Terminal nodes			Splitting: own lags (%)	
Model	Mean no.	Max. no.	Max. Obs.	Initial	Other
BART homosk.	2.33 (0.020)	5.23 (0.294)	165.39 (0.963)	0.19 (0.011)	0.20 (0.018)
mixBART homosk.	2.33 (0.021)	5.28 (0.297)	165.44 (0.943)	0.19 (0.011)	0.20 (0.018)
errorBART homosk.	2.21 (0.022)	4.80 (0.267)	165.92 (1.084)	0.18 (0.011)	0.20 (0.023)
fullBART homosk.	2.31 (0.018)	5.17 (0.285)	165.59 (0.882)	0.18 (0.012)	0.19 (0.017)
BART SV	2.33 (0.019)	5.25 (0.284)	165.59 (0.885)	0.19 (0.010)	0.20 (0.019)
mixBART SV	2.33 (0.021)	5.29 (0.301)	165.43 (0.979)	0.19 (0.011)	0.20 (0.018)
errorBART SV	2.21 (0.022)	4.82 (0.267)	166.02 (0.960)	0.18 (0.011)	0.20 (0.021)
fullBART SV	2.32 (0.021)	5.18 (0.308)	165.36 (0.997)	0.18 (0.010)	0.19 (0.018)
BART hBART	2.33 (0.021)	5.26 (0.284)	165.22 (0.919)	0.19 (0.011)	0.20 (0.018)
mixBART hBART	2.33 (0.021)	5.27 (0.305)	165.12 (1.016)	0.19 (0.011)	0.20 (0.019)
errorBART hBART	2.22 (0.021)	4.80 (0.271)	165.90 (0.931)	0.18 (0.010)	0.20 (0.021)
fullBART hBART	2.32 (0.021)	5.18 (0.274)	165.54 (1.015)	0.18 (0.011)	0.19 (0.018)

Notes: Mean and standard error (in parentheses) over 50 replications of the DGPs.

Table D.2: Tree diagnostics for the small TVP DGPs.

	Terminal nodes			Splitting: own lags (%)	
Model	Mean no.	Max. no.	Max. Obs.	Initial	Other
BART homosk.	2.32 (0.020)	5.22 (0.277)	165.22 (0.968)	0.19 (0.010)	0.20 (0.019)
mixBART homosk.	2.32 (0.020)	5.22 (0.263)	165.14 (0.905)	0.19 (0.012)	0.20 (0.019)
errorBART homosk.	2.22 (0.022)	4.81 (0.285)	166.62 (0.975)	0.18 (0.011)	0.20 (0.020)
fullBART homosk.	2.31 (0.019)	5.16 (0.274)	165.27 (0.872)	0.18 (0.011)	0.19 (0.018)
BART SV	2.33 (0.020)	5.24 (0.276)	165.10 (0.927)	0.19 (0.011)	0.20 (0.019)
mixBART SV	2.32 (0.020)	5.24 (0.284)	165.06 (0.870)	0.19 (0.011)	0.20 (0.018)
errorBART SV	2.22 (0.021)	4.83 (0.281)	166.67 (0.919)	0.18 (0.011)	0.20 (0.022)
fullBART SV	2.31 (0.019)	5.14 (0.294)	165.30 (0.938)	0.18 (0.011)	0.19 (0.018)
BART hBART	2.33 (0.020)	5.27 (0.279)	165.05 (0.919)	0.19 (0.011)	0.20 (0.018)
mixBART hBART	2.33 (0.021)	5.25 (0.302)	165.10 (1.037)	0.19 (0.011)	0.20 (0.018)
errorBART hBART	2.22 (0.020)	4.83 (0.288)	166.69 (0.936)	0.18 (0.011)	0.20 (0.020)
fullBART hBART	2.32 (0.021)	5.18 (0.291)	165.28 (0.950)	0.18 (0.010)	0.19 (0.018)

Notes: Mean and standard error (in parentheses) over 50 replications of the DGPs.

Table D.3: Tree diagnostics for the large constant parameter DGPs.

	Terminal nodes			Splitting: own lags (%)	
Model	Mean no.	Max. no.	Max. Obs.	Initial	Other
BART homosk.	2.33 (0.009)	5.25 (0.124)	164.54 (0.364)	0.18 (0.005)	0.19 (0.008)
mixBART homosk.	2.34 (0.010)	5.26 (0.126)	164.54 (0.419)	0.18 (0.004)	0.19 (0.008)
errorBART homosk.	2.12 (0.010)	4.50 (0.117)	168.56 (0.468)	0.16 (0.005)	0.18 (0.011)
fullBART homosk.	2.32 (0.009)	5.18 (0.130)	164.52 (0.388)	0.16 (0.005)	0.17 (0.008)
BART SV	2.33 (0.009)	5.25 (0.140)	164.68 (0.409)	0.18 (0.005)	0.19 (0.008)
mixBART SV	2.33 (0.010)	5.24 (0.123)	164.68 (0.439)	0.18 (0.005)	0.18 (0.008)
errorBART SV	2.13 (0.010)	4.51 (0.109)	168.60 (0.496)	0.16 (0.005)	0.18 (0.009)
fullBART SV	2.32 (0.009)	5.19 (0.127)	164.67 (0.424)	0.16 (0.005)	0.17 (0.008)
BART hBART	2.34 (0.009)	5.28 (0.140)	164.37 (0.386)	0.18 (0.005)	0.18 (0.008)
mixBART hBART	2.34 (0.009)	5.27 (0.133)	164.30 (0.400)	0.18 (0.005)	0.18 (0.007)
errorBART hBART	2.12 (0.010)	4.49 (0.129)	168.74 (0.467)	0.16 (0.004)	0.18 (0.010)
fullBART hBART	2.33 (0.009)	5.22 (0.128)	164.62 (0.418)	0.16 (0.005)	0.17 (0.007)

Notes: Mean and standard error (in parentheses) over 50 replications of the DGPs.

Table D.4: Tree diagnostics for the large TVP DGPs.

	Terminal nodes			Splitting: own lags (%)	
Model	Mean no.	Max. no.	Max. Obs.	Initial	Other
BART homosk.	2.33 (0.010)	5.23 (0.127)	165.18 (0.432)	0.18 (0.005)	0.18 (0.008)
mixBART homosk.	2.33 (0.009)	5.23 (0.139)	165.17 (0.419)	0.17 (0.005)	0.19 (0.008)
errorBART homosk.	2.11 (0.010)	4.45 (0.111)	168.93 (0.459)	0.16 (0.005)	0.18 (0.010)
fullBART homosk.	2.32 (0.009)	5.16 (0.126)	165.16 (0.415)	0.16 (0.005)	0.17 (0.008)
BART SV	2.33 (0.010)	5.23 (0.127)	165.24 (0.432)	0.18 (0.005)	0.18 (0.008)
mixBART SV	2.33 (0.009)	5.23 (0.135)	165.30 (0.432)	0.17 (0.005)	0.18 (0.008)
errorBART SV	2.11 (0.010)	4.46 (0.119)	169.04 (0.456)	0.16 (0.004)	0.18 (0.011)
fullBART SV	2.32 (0.009)	5.18 (0.129)	165.22 (0.385)	0.16 (0.005)	0.17 (0.008)
BART hBART	2.34 (0.009)	5.25 (0.126)	164.96 (0.403)	0.18 (0.005)	0.18 (0.008)
mixBART hBART	2.34 (0.009)	5.26 (0.127)	164.92 (0.416)	0.18 (0.004)	0.18 (0.008)
errorBART hBART	2.11 (0.010)	4.44 (0.108)	169.14 (0.441)	0.16 (0.005)	0.18 (0.011)
fullBART hBART	2.33 (0.010)	5.21 (0.127)	165.20 (0.393)	0.16 (0.005)	0.17 (0.008)

Notes: Mean and standard error (in parentheses) over 50 replications of the DGPs.