Ehmann, Daniel

Working Paper

Dynamic Factor Trees and Forests: A theory-led machine learning framework for non-linear and state-dependent short-term U.S. GDP growth predictions

KOF Working Papers, No. 472

Provided in Cooperation with:
KOF Swiss Economic Institute, ETH Zurich

Suggested Citation: Ehmann, Daniel (2022) : Dynamic Factor Trees and Forests: A theory-led machine learning framework for non-linear and state-dependent short-term U.S. GDP growth predictions, KOF Working Papers, No. 472, ETH Zurich, KOF Swiss Economic Institute, Zurich, https://doi.org/10.3929/ethz-b-000544472

This Version is available at:
https://hdl.handle.net/10419/258982

Dynamic Factor Trees and Forests - A Theory-led Machine Learning Framework for Non-Linear and State-Dependent Short-Term U.S. GDP Growth Predictions

ETH zürich

KOF Swiss Economic Institute

Dynamic Factor Trees and Forests

A Theory-led Machine Learning Framework for Non-Linear and State-Dependent Short-Term U.S. GDP Growth Predictions

Daniel Ehmann

KOF Working Papers, No. 472, updated version, April 2022

KOF

Dynamic Factor Trees and Forests

A Theory-led Machine Learning Framework for Non-Linear and State-Dependent Short-Term U.S. GDP Growth Predictions

Daniel Ehmann

ETH Zurich

Abstract

This article proposes so-called Dynamic Factor Forests (DFF) for macroeconomic forecasting, which extend the model-based *trees* of Zeileis, Hothorn, and Hornik (2008) in the spirit of Garge, Bobashev, and Eggleston (2013) to model-based *forests* and synthesize the recent machine learning, business cycle and dynamic factor model literature within a unified statistical machine learning framework. DFFs correspond to state-dependent, non-linear and smoothed forecasting models that allow us to embed theory-led factor models in powerful tree-based machine learning ensembles conditional on, for instance, the business cycle state. An extensive out-of-sample forecasting experiment for short-term U.S. GDP growth provides encouraging results in favor of DFFs.

1. Introduction[a]

Machine learning systems in general and deep learning systems in particular have achieved notable breakthroughs in predictive accuracy in recent years (cf. e.g. Esteva et al., 2017; LeCun, Bengio, & Hinton, 2015; McAfee & Brynjolfsson, 2017). Unfortunately, however, the predictive ability and interpretative feasibility of these systems stand usually in conflict with one another (Breiman, 2001b) as their inner working mechanisms are typically complex and opaque and therefore difficult to understand — an issue known as the "black box" problem (Appenzeller, 2017; Mittelstadt et al., 2016; Mittelstadt & Floridi, 2016; Mukherjee, 2017; Veltri, 2017).

As the algorithmic modeling principles embedded in most machine learning algorithms are fundamentally different from the stochastic modelling principles employed in conventional econometric techniques, the two paradigms give rise to different strengths: While the former typically excels at predictive tasks, the latter is particularly strong at inferential analyses (cf. Athey & Imbens, 2019; Breiman, 2001b; McAfee & Brynjolfsson, 2017; Mullainathan & Spiess, 2017; Veltri, 2017; Wochner, 2018). A promising new field of research in *statistical* machine learning seeks to fruitfully merge these two paradigms so as to endow machine learning algorithms with theory-led and parametric models while maintaining their high predictive accuracy (cf. Athey & Imbens, 2017, 2019; Seibold et al., 2016; Varian, 2014; Zeileis et al., 2008). To take an example outside the field of economics – which inspired this research – consider the benefits of treatment effects in medical applications, which are likely to differ between individuals and may depend, for instance, on their demographics and health conditions (cf. Athey & Imbens, 2006, 2017, p. 10; Seibold et al., 2016). Seibold et al. (2016) promote advances in personalized medical healthcare by employing a model-based tree that is able to autonomously identify distinct patient subgroups with heterogenous medical treatment effects based on a set of personal health conditions.

This article adopts and adapts these ideas for macroeconomic forecasting and proposes so-called Dynamic Factor Forests (DFF), which extend model-based *trees* proposed in Zeileis et al. (2008) to model-based *forests* in the spirit of Garge et al. (2013) and synthesize the recent machine learning, dynamic factor model and business cycle literature within a unified statistical machine learning framework for model-based (MOB) recursive partitioning. DFFs are non-linear, state-dependent and smoothed forecasting models, which reduce to the standard Dynamic Factor Model (DFM) as a special case and allow us to embed theory-led factor models in powerful tree-based machine learning ensembles conditional on, for instance, the state of the business cycle. A dynamic factor forest combines hundreds of dynamic factor trees (DFT), each of which is grown from a (block-) bootstrapped sample (cf. Garge et al., 2013; see Breiman, 1996a, 2001a; Hastie, Tibshirani, & Friedman, 2009, for bootstrap aggregation (bagging) of (decorrelated) trees; see Canty, 2002; Politis & Romano, 1994, for block-bootstrapping).

[a] A previous version of this paper was published under the same title in the KOF Working Paper Series of ETH Zurich in 2020 as well as under the title "Dynamic Factor Forests – A Theory-led Machine Learning Framework for Non-Linear and State-Dependent Short-Term U.S. GDP Growth Predictions" in Ehmann (2021) (see acknowledgements).

Dynamic factor trees therefore consist of two key ingredients (cf. Zeileis et al., 2008): First, to benefit from big macroeconomic datasets, we build upon the recent dynamic factor model literature (e.g. Stock & Watson, 2011, 2016; Siliverstovs & Wochner, 2021) and embed factor-augmented autoregressive processes as parametric models within the MOB-framework. While we begin with traditional factors derived from principal components, our framework adopts a deliberately broad notion of factors as we also examine novel kinds of factor extraction methods in the spirit of the recent forecasting literature (Ehmann, 2020; H. H. Kim & Swanson, 2018; Tu & Lee, 2019). Second, analogous to the health conditions of patients in Seibold et al. (2016), we choose, in the spirit of the business cycle literature (e.g. Chauvet & Potter, 2013; Doz & Fuleky, 2020, for reviews), the health conditions of the economy (proxied via recession probability indices) as partitioning variable to derive the model-based regression tree structure. While we start with recession probability indices as a single partitioning variable, the conditioning sets are subsequently extended to more complex ones in order to allow for generalized and time-varying state-dependent patterns as the MOB-framework autonomously detects interactions among the partitioning variables in the conditioning set (cf. e.g. Zeileis et al., 2008; Goulet Coulombe, 2020). A dynamic factor forest is then estimated via Zeileis et al.'s (2008) and Garge et al.'s (2013) model-based recursive partitioning algorithm in six steps: (1) a dynamic factor model is estimated, (2) parameter instability tests with regards to (a random subset of) the partitioning variable(s) are performed, (3) if instabilities are present, the initial parametric model is autonomously split in a tree-like fashion into two sub-states such that the local fit in each sub-state is maximized, (4) the procedure is recursively repeated in each sub-state until a stopping criterion (such as minimum state size) is reached (also see Zeileis & Hothorn, 2015; Hothorn & Zeileis, 2020). (5) As deeply grown trees are likely to overfit, we consider in addition to pruning strategies (cf. e.g. Zeileis & Hothorn, 2015; Zeileis et al., 2008; James, Witten, Hastie, & Tibshirani, 2013), the inclusion of Ridge- and Lasso-based regularization techniques to enable a state-dependent regularization (cf. Quan, Wang, Gan, & Valdez, 2020). (6) Steps (1) to (5) return a single DFT and DFFs are obtained in analogy to Breiman's (1996a) bagging by repeating these steps for a large number of bootstrapped samples and aggregating their predictions (cf. Garge et al., 2013). While each DFT typically results in multiple regimes with a sharp transition between regimes (cf. Zeileis et al., 2008), the bagging principles in DFFs allow to turn these hard into soft transitions such that DFFs essentially represent a smoothed version of DFTs (cf. Bühlmann & Yu, 2002; Schlosser et al., 2019).

Based on Stock and Watson's (2017) views in their summary article on time series econometrics, we see both conceptual and empirical advantages to motivate such a modeling strategy. On the conceptual side, Stock and Watson (2017) expect that "the next steps towards exploiting additional information in large datasets will need to use new statistical methods guided by economic theory." (ibid., p. 83). Model-based recursive partitioning may provide a promising framework in this regard (cf. Veltri, 2017; Zeileis et al., 2008, for a general dis-

cussion): While standard regression tree algorithms split the feature space into non-overlapping subgroups and fit a trivial model, such as the sample mean, to each of these (cf. e.g. Breiman, Friedman, Stone, & Olshen, 1984; Hastie et al., 2009; James et al., 2013), dynamic factor trees and forests, instead, allow us to fit non-trivial dynamic factor models to each subgroup (e.g. Zeileis et al., 2008), which not only deliver great empirical performance but are also well-grounded in macroeconomic equilibrium theories (cf. e.g. Forni, Giannone, Lippi, & Reichlin, 2009; Stock & Watson, 2006, 2011, 2016, 2017, and references therein; also e.g. Diebold & Rudebusch, 1996, p. 69ff.).[1] On the empirical side, Stock and Watson (2017) refer to the forecasting models' repeated inability to capture the severity of economic downturns as the "Mother of All Forecast Errors" (ibid., p. 82) and question in the context of dynamic factor models "whether there is exploitable nonlinear structure [...] that could perhaps be revealed by modern machine learning methods." (ibid., p. 83). The proposed state-dependent machine learning approach may contribute along these lines in that they allow for non-linearities by fitting distinct models to autonomously detected subsets of the data (cf. Seibold et al., 2016; Zeileis et al., 2008) and thereby equip the models (in analogy to the regime switching literature) with the flexibility to react differently to movements of the underlying series in case of deteriorating and improving economic conditions (cf. e.g. K. Kim & Swanson, 2016; also Doz & Fuleky, 2020, for a review). Siliverstovs and Wochner's (2021) results motivate such a state-dependent modeling design too.

This article contributes to the current state of the literature by synthesizing three burgeoning streams of the macroeconomic forecasting literature within a unified statistical machine learning framework in a mixed-frequency setup and several points deserve consideration: First, the machine learning literature put forth an increasing number of studies reporting that standard tree-based ensembles, such as random forests, can have great predictive power for a range of macroeconomic and financial indicators (e.g. J. C. Chen et al., 2019; Ehmann, 2020; Goulet Coulombe et al., 2019; Khaidem et al., 2016; Medeiros et al., 2021; Wochner, 2018; also see Garcia, Medeiros, & Vasconcelos, 2017 who find satisfactory performance). While standard regression trees belong to the class of non-parametric methods (cf. X. Chen & Ishwaran, 2012), model-based trees are closely related yet clearly distinct in that they embed parametric models within tree-based structures and are thus semi-parametric in nature (Jones, Mair, Simon, & Zeileis, 2020; Zeileis et al., 2008). Second, the proposed modeling approach is closely related to the business cycle literature that combines regime-switching ideas with dynamic factor models so as to treat expansions and recessions as distinct stochastic entities (Diebold & Rudebusch, 1996; see discussions in Diebold, 2003 and Doz & Fuleky, 2020, and references therein). For instance, Chauvet and Potter (2013) show in their comprehensive assessment of leading forecasting models that Markov-Switching Dynamic Factor Models (MS-DFMs) rank among the best models in terms of out-of-sample forecasting performance, especially during recessions. While these MS-DFMs and their recent extensions (e.g. Camacho et al., 2018; Doz et al., 2020) may attain good predictive performance and

accurately replicate empirically observed facts, they are limited in at least two respects: On the one hand, they are often constrained to switch between two states, which may, according to Leiva-Leon et al. (2020) and Carstensen et al. (2020), be too restrictive because the depth and severity of recessions is likely to differ across the sample and precludes the models' ability to account for more subtle state-dependent patterns. On the other hand, any endeavours to allow for more than a selected few states in Markov-Switching models can quickly result in serious computational problems (Audrino, 2006). Model-based recursive partitioning principles employed in our DFTs and DFFs, instead, are less subject to these restrictions: They do not fix the number of states à-priori but determine the number of states endogenously and are able to autonomously adapt the number of states over time (cf. Zeileis et al., 2008; also Audrino, 2006). At the same time, DFTs are computationally fairly efficient and the computational costs for DFFs scale, in analogy to regression trees and forests, approximately linearly in the number of DFTs and may thus be weighed against the desired level of accuracy or stability (cf. Biau & Scornet, 2016). Third, our dynamic factor forests are related to threshold models in general and smooth transition regression (STR) models in particular (see e.g. Kock & Teräsvirta, 2011, for a review). While STR models rely on an exogenous choice of a parametric transition function to smoothen the transition between states (Kock & Teräsvirta, 2011, and references therein), the bootstrap aggregation principles embedded in dynamic factor forests offer not only the advantage to smoothen the decision boundaries in a non-parametric and purely data driven way but can also accommodate smoothing over multiple regimes (cf. Bühlmann & Yu, 2002; Schlosser et al., 2019). The idea to bridge parametric threshold models and tree-based machine learning algorithms is related to previous work in Da Rosa et al. (2008) who embed STRs in non-parametric regression trees as well as Audrino (2006) and Audrino and Medeiros (2011) who fit parametric AR-GARCH models to local, tree-structured partitions of the data and thereby carefully incorporate the dynamics in the conditional variance for their short-term predictions of the interest rate. Concerning the specification of the conditional mean, Audrino and co-authors fit autoregressive (GARCH-type) processes to tree-based partitions whereas we augment the local models with information from a vast number of predictors and fit factor-augmented autoregressive processes to tree-structured partitions of the data. Fourth, as indicated above, while we start with a deliberately narrowly specified conditioning set, we subsequently extend to more complex variations to account for possibly generalized and time-varying state-dependent dynamics (e.g. Goulet Coulombe, 2020 and Zeileis et al., 2008, and references therein; also Audrino & Medeiros, 2011). A valuable property of DFFs in case of such multivariate conditioning sets is the randomized variable selection during tree estimation in analogy to Breiman's (2001a) random forest, which allows to decorrelate the different DFTs for enhanced predictive performance (cf. Hastie et al., 2009, James et al., 2013, Zeileis et al., 2008, Hothorn & Zeileis, 2020, Garge et al., 2013, and references therein). Fifth, building upon Quan et al. (2020), we extend the framework of Zeileis et al. (2008) from OLS-based to

Lasso- and Ridge-based state-dependent model fitting in the terminal nodes of deeply grown (dynamic factor) trees. This inclusion of state-dependent Ridge and Lasso penalization (see Hastie et al., 2009, for a review) offers an alternative regularization approach as opposed to the common tree-based pruning techniques (Zeileis & Hothorn, 2015; Zeileis et al., 2008, for model-based trees; Breiman et al., 1984; James et al., 2013, for conventional trees). We consider this extension as particularly interesting because it not only allows to mitigate overfitting but also enables the different state-dependent models to adopt conceptually distinct model designs in a data-driven way (cf. James et al., 2013). For instance, we could think of a situation where the state-dependent Lasso regularization imposes an autoregressive process in expansionary states and a factor augmented autoregressive process in contractionary ones (cf. K. Kim & Swanson, 2016). Indeed, we find that this type of regularization can perform superior to pruning mechanisms and yields among the very best results when combined with conditioning sets that allow for time-varying state-dependent dynamics. The generalized time-varying parameter (GTVP) model of Goulet Coulombe (2020) is a closely related, interesting variation that allows to fit parametric models to tree-like partitions of the data. GTVPs grow all trees deep and impose Ridge as well as random walk restrictions to regulate GTVPs. Sixth, we contribute to the literature by examining the impact of using novel instead of traditional PCA-based factors within DFFs in the spirit of the most recent forecasting literature (Ehmann, 2020; H. H. Kim & Swanson, 2018; Tu & Lee, 2019) and examine the merits of Kernel and Sparse Principal Components as well as Partial Least Squares (Hastie et al., 2009, for a review). Similar to these studies, we shall also consider both plain and targeted versions of the factors (Bai & Ng, 2008). Finally, as discussed above, we start with recession probability indices as a single partitioning variable that we obtain from external sources (see Section 3 for details). In this sense, the approach outlined heretofore can be considered as a single-stage machine learning approach. However, there is a growing literature that strives to predict recession probabilities more accurately in data-rich environments by means of machine learning techniques (e.g. Döpke, Fritsche, & Pierdzioch, 2017; Ng, 2014; Pierdzioch & Gupta, 2020; also see Doz & Fuleky, 2020 for a recent discussion). We build upon these advances and expand the analyses to a two-stage machine learning framework, where we predict the probability of a recession in a first stage and use it subsequently (possibly in combination with generalized and penalized extensions outlined above) within the proposed DFTs and DFFs as a second stage. Our approach to predict recession probabilities is closely related to the ones adopted in, for instance, Ng (2014), Döpke et al. (2017) and Pierdzioch and Gupta (2020) but distinct in that we additionally implement a smoothing step. Smoothing is inspired by Ng (2014) who expressed a lack of sufficient persistence in her machine learning-based recession probability predictions as well as Chauvet and Hamilton (2006) and Hamilton (2018) who find smoothened recession probabilities to be useful for inferring the likely future state of the economy.

This article assesses the empirical performance of the proposed dynamic factor trees and forests against their two constituent benchmarks, standard dynamic

factor models and standard regression trees and forests, as well as conventional benchmarks, such as autoregressive and distributed lag processes. The main analyses are based on three distinct FRED datasets (FRED-MD, FRED-QD, FRED; see Section 3), which can be combined to a balanced panel with real GDP growth as dependent variable, 375 explanatory variables plus a recession probability indicator (RPI) from October 1967 to September 2018 (e.g. FRED, 2019; McCracken, 2019; McCracken & Ng, 2016, 2019a, 2019b, 2020). The analyses make two key modeling assumptions: First, to cope with mixed frequencies, we follow Kim and Swanson's (2014) approach and interpolate quarterly GDP to a monthly frequency. We will show the robustness of our results by means of two alternative interpolation methods (see Section 4.3.1) and run the experiment also in quarterly frequency (see Section 4.3.4). Second, as a consequence, we also interpolate the recession probability index to a monthly frequency and assume the series to be released together with all other monthly indicators in FRED-MD. Under this interpolation and publication scheme, we consider to have an accurate monthly proxy for the RPI available. This assumption also allows us to make our results more comparable with the existing Markov-switching literature, which typically derives the state probabilities from the most recent set of available predictors (cf. Kuan, 2002, and references therein). We assess the sensitivity of this assumption by predicting the missing values on the current edge (see Section 4.3.8) and show robustness for two alternative RPIs (see Section 4.3.7).

We find considerable empirical evidence in favor of the proposed dynamic factor trees and forests. Our out-of-sample forecasting experiment shows that they yield significant gains over standard dynamic factor models and show that the extension from DFTs to DFFs yields valuable improvements in predictive accuracy. Interestingly, our evaluations à la Chauvet and Potter (2013) and Siliverstovs and Wochner (2021) show that DFFs state-dependent model design allows to systematically improve upon DFMs in both expansionary and recessionary periods. In view of the generally strong performance of DFMs in settings with many predictors (cf. Chauvet & Potter, 2013; Stock & Watson, 2017), we perceive these additional improvements as notable gains in predictive accuracy. Additionally, we find carefully designed narrow conditioning sets to perform generally superior to generic and broad sets that simply include a vast number of features. Nonetheless, even in the latter case, the model is typically still able to outperform standard dynamic factor models. With regards to model tuning, we find pruning and penalization strategies (Ridge and Lasso) to work both generally well. Yet, we also show that one can be clearly better: For instance, in case of DFTs, Lasso tends to perform persistently better than Ridge and quite frequently also (slightly) better than pruning strategies. Furthermore, the novel factors tend to show generally good performance but they are not substantially superior or inferior to PCA-based factors, which is broadly consistent with the findings reported in Ehmann (2020). Interestingly, we also find for these novel types in the present framework that a targeting step as in Bai and Ng (2008) typically results in superior performance. On top of that, we also document that the extension to a two-stage machine learning framework can be successful. Specifically,

we find that the internalization of recession probability predictions via nonlinear and regularized linear machine learners and smoothers can result in DFFs that outperform dynamic factor models in statistically significant ways (especially for one-step ahead predictions). Finally, our findings qualify as fairly robust against common robustness tests, which range from alternative interpolation and aggregation methods to extended evaluation windows and rolling estimation schemes over to alternative recession probability indices and model specifications.

Building upon Siliverstovs and Wochner (2021), Ehmann (2020) as well as Wochner (2018), the remainder of this article is organized as follows: Section 2 sets out the modeling environment and formalizes dynamic factor trees and forests. Section 3 describes the data. Section 4 evaluates the models' forecasting performance and their robustness. Section 5 concludes with directions for future research.

2. Modeling Framework

2.1. Setup, Notation and Environment

For the definition of our modeling framework, we follow the notational and methodical conventions in the relevant literature (e.g. Elliott & Timmermann, 2016; Stock & Watson, 2006, 2016; Siliverstovs & Wochner, 2021; Wochner, 2018). The timeline $t \in \{1, \dots, T\}$ is divided into an estimation window (Dec. 1967 – Dec. 1997) and forecasting window (Jan. 1998 – Sep. 2018) and let Q denote the last time period of the first estimation window (Dec. 1997), so that the recursive estimation window is given as $\mathscr{S}_e = \{1, \dots, \tau\}$ with $\tau \in \{Q, \dots, T-h\}$, and the forecasting window as $\mathscr{S}_f = \{Q+1, \dots, T-h+1\}$, where $h \in \{1, 3\}$ corresponds to the monthly forecasting horizon (cf. e.g. Siliverstovs & Wochner, 2021).

Denoting vectors and matrices in bold letters, the dataset consists of four different types of stationary variables: The dependent variable, $Y_t^{(h)} \in \mathbb{R}$, the set of K mean-zero and unit-variance standardized explanatory variables, $\boldsymbol{X}_t^{(1)} \in \mathbb{R}^K$ (including their 1st and 2nd-order lags; cf. H. H. Kim & Swanson, 2014) as well as the partitioning variables, $\boldsymbol{Z}_t^{(1)} \in \mathbb{R}^P$, which are used in the MOB-framework to segment the sample space into S distinct subsets (cf. Zeileis et al., 2008). We extract the factors, $\boldsymbol{F}_t^{(1)} \in \mathbb{R}^R$, analogously to Stock and Watson (2009) and Siliverstovs and Wochner (2021) from the full sample via principal components analysis (PCA) (e.g. Stock & Watson, 2002b, 2006, 2016, for details). These full factor estimates are stable and can be consistently estimated provided that instabilities are limited (see Bates, Plagborg-Møller, Stock, & Watson, 2013, for precise definitions; also Bai & Han, 2016 and Stock & Watson, 2002a, 2016, and references therein). Moreover, besides these plain factors extracted from all predictors, we consider targeted factors in the spirit of Bair et al. (2006) and Bai and Ng (2008) derived from a targeted set of predictors via hard-thresholding at the 5% level when considering the pairwise correlations between predictors and the response. As outlined above, in addition to plain and targeted PCA-based factors above, we follow the recent literature (e.g. Ehmann, 2020; H. H. Kim &

Swanson, 2018; Tu & Lee, 2019) and examine plain and targeted factors derived from Partial Least Squares (PLS) of Wold (1975 as cit. in Hastie et al., 2009, p. 94), Kernel Principal Component Analysis (KPCA) proposed in Schölkopf et al. (1997) as well as factors derived from Sparse Principal Component Analysis (SPCA) proposed in Zou et al. (2006) from all as well as a targeted subset of variables.[2] As summarized in Ehmann (2020), PLS extracts factors in a supervised fashion (also see Hastie et al., 2009), KPCA corresponds to a non-linear extension of PCA (also see Schölkopf et al., 1997) whereas SPCA method introduces sparse factor loadings for better interpretability and mitigation of overfitting in high-dimensional settings (also see Erichson et al., 2020; Zou et al., 2006). We refer the reader to Ehmann (2020) for a summary of the methods and to Hastie et al. (2009) for a comprehensive treatment. Further, the bracketed super-indices indicate the h-period ahead stationarity transformation of dependent and independent variables (cf. e.g. Stock & Watson, 2012). Finally, let $\boldsymbol{W}_t^{(1)} = \left(1, Y_t^{(1)}, \ldots, Y_{t-L+1}^{(1)}\right)'$ denote additional controls that include an intercept constant and autoregressive dependent variables up to lag L, such that the relevant dataset for model estimation is given as $\boldsymbol{D}_t^{(h)} = \left(Y_{t+h}^{(h)}, \boldsymbol{I}_t^{(1)\prime}\right) = \left(Y_{t+h}^{(h)}, \boldsymbol{W}_t^{(1)\prime}, \boldsymbol{F}_t^{(1)\prime}, \boldsymbol{Z}_t^{(1)\prime}\right)$. The general form of the predictive model can then be given as,

$$Y_{t+h}^{(h)} = \mu_{t+h}^{(h)} + \epsilon_{t+h}^{(h)} = f^{(h)}\left(\boldsymbol{I}_t^{(1)}\right) + \epsilon_{t+h}^{(h)}, \tag{1}$$

where $Y_{t+h}^{(h)}$ is the dependent variable in period $t+h$, $f^{(h)}\left(\boldsymbol{I}_t^{(1)}\right)$ some horizon-specific predictive function of the inputs $\boldsymbol{I}_t^{(1)}$ and $\epsilon_{t+h}^{(h)}$ the error term (e.g. Hastie et al., 2009; James et al., 2013; Medeiros et al., 2021).

2.2. Dynamic Factor Trees[b]

2.2.1. Formal Model

Simply put, Zeileis et al.'s (2008) model-based recursive partitioning algorithm is a general statistical framework that allows to fit parametric models to distinct subsets of the data. As the authors explain, a key advantage of their framework is that these subsets (or states) are autonomously detected through the recursive application of parameter instability tests over the space spanned by the partitioning variable(s) in a tree-based fashion. Their algorithm is therefore capable to autonomously detect nonlinearities arising from interactions of variables and fits, in case of their presence, local models to subsets of the data that yield better fit than a single global model for all observations (cf. Zeileis & Hothorn, 2015; Zeileis et al., 2008). Their implementation therefore requires the specification of two key ingredients: The parametric model and partitioning variable. With regards to the former we build upon the recent factor model literature (e.g. Stock & Watson, 2011, 2016; Siliverstovs & Wochner, 2021) and employ factor-augmented autoregressive processes. With regards to the latter, we use in line with the business cycle and structural breaks literature (Chauvet & Potter, 2013, and Doz &

[b] As indicated in the text, parts of this section follow closely Zeileis et al. (2008) as well as Zeileis and Hothorn (2015) (also see Kopf, Augustin, & Strobl, 2013, for a thorough review).

Fuleky, 2020, and references therein), recession probability indices as partitioning variable within the MOB-framework.

To formalize their approach, let $\mathcal{M}(\boldsymbol{\psi}_s, \boldsymbol{\phi}_s; \boldsymbol{D}^{(h)})$ denote the parametric dynamic factor model with parameter coefficients $\boldsymbol{\psi}_s$ and $\boldsymbol{\phi}_s$ in state $s \in \{1, \dots, S\}$ for dataset $\boldsymbol{D}^{(h)}$, the MOB algorithm seeks to minimize the objective function, $\sum_{t \in \mathscr{S}_e} \Omega(\boldsymbol{\psi}_s, \boldsymbol{\phi}_s; \boldsymbol{D}_t^{(h)})$, for a given loss function $\Omega(\cdot)$ (Zeileis et al., 2008). As detailed in Zeileis et al. (2008), Zeileis and Hothorn (2015) and Kopf et al. (2013), the algorithm starts in the first state with all observations and derives a state-dependent model in case of a single partitioning variable in four key steps:

(1) estimate the parametric model, $\mathcal{M}(\cdot)$, in the current state by minimizing the objective function (e.g. via OLS if $\Omega(\cdot)$ is the squared residual loss);

(2) run score-based fluctuation tests for parameter instability over all P partitioning variables in $\boldsymbol{Z}_t$ and if instabilities are present (at significance level α), determine the variable $p^\star$ with the most significant instability; else, stop;

(3) determine the optimal split point $\zeta^\star$ for $Z_{p^\star,t}^{(1)}$ such that the objective function, given as the sum of local objective functions in the two resulting sub-states, $\sum_{t \in \{t | Z_{p^\star,t} \leq \zeta\}} \Omega(\boldsymbol{\psi}_{s_1}, \boldsymbol{\phi}_{s_1}; \boldsymbol{D}_t^{(h)}) + \sum_{t \in \{t | Z_{p^\star,t} > \zeta\}} \Omega(\boldsymbol{\psi}_{s_2}, \boldsymbol{\phi}_{s_2}; \boldsymbol{D}_t^{(h)})$, is minimized;

(4) split the current state according to $(p^\star, \zeta^\star)$ into two sub-states, namely $\mathscr{S}_e^{(s_1)} = \{t | Z_{p^\star,t} \leq \zeta^\star\}$ and $\mathscr{S}_e^{(s_2)} = \{t | Z_{p^\star,t} > \zeta^\star\}$, and repeat the procedure in each substate until stability is achieved or the minimum node size, η, is reached.

A few points deserve further attention: First, the partitioning variable is allowed to be either be categorical or numerical, where the former provides the possibility to exogenously determine different states (via manual discretization of a numeric variable) whereas the latter allows to endogenously determine different states (based on automated detection mechanisms) (cf. Strobl, Wickelmaier, & Zeileis, 2011; Zeileis & Hornik, 2007; Zeileis et al., 2008). Second, the empirical parameter instability tests are based on generalized M-fluctuation tests that assess whether or not the scores of estimated objective functions $\hat{\boldsymbol{\omega}}_{s,t} = \partial\big(\Omega(\hat{\boldsymbol{\psi}}_s, \hat{\boldsymbol{\phi}}_s; \boldsymbol{D}_t^{(h)})\big) / \partial\big((\hat{\boldsymbol{\psi}}_s', \hat{\boldsymbol{\phi}}_s')\big)$ deviate systematically from zero over $Z_{p^\star,t}^{(1)}$ and use robust standard errors to account for possibly heteroscedastic and autocorrelated errors (Zeileis et al., 2008, p. 496ff. and Zeileis & Hornik, 2007, and references therein). These fluctuation or instability tests comprise several well-known tests employed in the macroeconomic literature on structural breaks, such as Andrews' (1993) Lagrange multiplier test for single discrete breaks in numeric variables (cf. Zeileis et al., 2008, p. 497f.; Rossi, 2013, p. 1236ff.; Zeileis, 2005), which were shown to be able to detect structural breaks in factor models (Breitung & Eickmeier, 2011) (see Zeileis et al., 2008, p. 496ff., and references therein for the case of categorical variables). Third, as the recursive testing procedure can entail multiple testing, the p-values are Bonferroni-adjusted (Zeileis & Hothorn, 2015, p. 4; Zeileis et al.,

2008). This procedure yields a non-linear and state-dependent dynamic factor model with S many states, which we call "Dynamic Factor Trees" (DFT). We may formalize these discussions as,

$$\mu_{t+h}^{(h)} = \sum_{s=1}^{S} \left(\boldsymbol{\psi}_s'(\lambda_s) \boldsymbol{W}_t^{(1)} + \boldsymbol{\phi}_s'(\lambda_s) \boldsymbol{F}_t^{(1)} \right) \mathbb{1}\left(\boldsymbol{Z}_t^{(1)} \in \mathcal{R}_s(\alpha, \gamma, \eta) \right) \tag{2}$$

where $\boldsymbol{\psi}_s$ and $\boldsymbol{\phi}_s$ denote the regression coefficients, $\mathcal{R}_s(\cdot)$ defines the set of observations in state s that depend (among others) on key hyper-parameters α, γ as well as η and $\mathbb{1}(\cdot)$ corresponds to the indicator function, which evaluates to unity whenever the condition is satisfied and to zero else (cf. e.g. Kock & Teräsvirta, 2011, and Chauvet & Potter, 2013 summarize closely related regime-switching models[3]; e.g. Breiman et al., 1984 and Hastie et al., 2009, p. 307ff. and p. 61ff., for conventional regression trees and penalized regression and e.g. Stock & Watson, 2002b, 2006, 2016, for DFMs; cf. Zeileis et al., 2008). Additionally, λ_s denotes the state-specific ℓ_1- or ℓ_2-norm penalty parameters arising from fitting Tibshirani's (1996) Lasso or Hoerl and Kennard's (1970) Ridge in the terminal nodes of each tree to shrink parameter coefficients towards zero (cf. Quan et al., 2020; Hastie et al., 2009). Moreover, notice that the dynamic factor model emerges as a special case from equation (2) in case of a single state $\mathcal{R}_1$ and the absence of penalization. Finally, equation (2) is related to Stock and Watson (2009) in that they find the best predictive model to be one that uses full sample factors and allows for state-dependent (or time-varying) regression coefficients.

2.2.2. Empirical Implementation

When trees are recursively grown at full depth, they tend to overfit the data, which can be mitigated via pruning (Breiman et al., 1984; James et al., 2013, p. 303ff.). The MOB framework offers two main tree pruning mechanisms to determine the optimal tree depth: A *pre*-pruning approach where statistical parameter instability tests serve as an early stopping criterion and prevent further tree growth when no significant instabilities can be detected (at significance level α) or *post*-pruning that initially grows the trees deep but subsequently prunes them back based on an information criterion, γ, such as BIC (e.g. Zeileis & Hothorn, 2015, and references therein; also Zeileis et al., 2008). While we examine pre-pruning mechanisms as a robustness test, our estimation results rely on post-pruning techniques, which are well-established in the tree-based literature (e.g. Audrino & Medeiros, 2011; James et al., 2013). Alternatively, another commonly applied procedure to optimize performance is hyper-parameter tuning (of the minimum state size, η) via cross-validation (CV) (cf. e.g. James et al., 2013). Following Wochner (2018), we implement Racine's (2000) blocked cross-validation (for η) that divides the sample into ten (approximately) equally long blocks and discards an equivalent of one year of observations on either side to mitigate dependencies between test and training sets. Finally, rather than fitting OLS to each terminal substate, we examine the impact of Ridge and Lasso penalization when trees are grown deep (cf. Quan et al., 2020), i.e. when neither

pruning nor minsize optimization is active.[4]

2.3. Dynamic Factor Forests

2.3.1. Formal Model

Our discussions above, have so far only considered a single dynamic factor tree. A single tree, however, is characterized by sharp (rather than smooth) transitions, which create instabilities and cause trees to be sensitive to minor changes in the data (Bühlmann & Yu, 2002, Breiman, 1996b, Garge et al., 2013, and references therein; James et al., 2013; Zeileis et al., 2008). Forests, in contrast, build upon bootstrap aggregation (bagging) principles and average among a multitude of trees, each of which is grown from bootstrapped samples (Breiman, 1996a, 2001a; Hastie et al., 2009; James et al., 2013), which alleviates the instabilities by turning the sharp thresholds of trees into soft thresholds, so that forests essentially represent a smoothed version of trees (cf. Bühlmann & Yu, 2002; Schlosser et al., 2019).

Garge et al. (2013) extended Zeileis et al.'s (2008) model-based recursive partitioning to tree ensembles. We extend dynamic factor trees in a similar vein and propose Dynamic Factor Forests (DFF) by augmenting the algorithmic procedure of DFTs with three steps: First, to decorrelate the individual DFTs analogously to Breiman's (2001a) random forests, we do not consider all P variables but rather a random subset of $\tilde{P} = \max(\lfloor P/3 \rfloor, 1)$ predictors in step 2 (e.g. Hastie et al., 2009, p. 587ff.; James et al., 2013; Garge et al., 2013; Zeileis et al., 2008; Hothorn & Zeileis, 2020). Second, following previous work (Ehmann, 2020; Medeiros et al., 2021; Wochner, 2018), each of the DFTs in DFFs is grown from a (stationary) block-bootstrapped sample (Politis & Romano, 1994; Canty, 2002, for details). Third, in the spirit of conventional bagging (cf. Breiman, 1996a; Hastie et al., 2009, p. 282ff.), we subsequently average among all fitted dynamic factor trees and may therefore characterize dynamic factor forests as,

$$\mu_{t+h}^{(h)} = \frac{1}{B} \sum_{b=1}^{B} \left[\sum_{s=1}^{S} \left(\boldsymbol{\psi}_{b,s}'(\lambda_{b,s}) \boldsymbol{W}_t^{(1)} + \boldsymbol{\phi}_{b,s}'(\lambda_{b,s}) \boldsymbol{F}_t^{(1)} \right) \mathbb{1}\left(\boldsymbol{Z}_t^{(1)} \in \mathcal{R}_{b,s}(\alpha, \gamma, \eta) \right) \right] \tag{3}$$

where the parameters $\boldsymbol{\psi}_{b,s}$ and $\boldsymbol{\phi}_{b,s}$ are the state-dependent parameters for the b-th bootstrap and $\mathcal{R}_{b,s}(\alpha, \gamma, \eta)$ denotes the s-th sub-state of the b-th dynamic factor tree.

2.3.2. Empirical Implementation

Our DFFs consist of $B = 500$ DFTs (unless mentioned otherwise), which appears to be sufficiently high given that Garge et al. (2013) propose $B = 300$ as default value for model-based forests. In analogy to DFTs, DFFs can apply a pruning or hyper-parameter tuning strategy. In the latter case, cross-validated η is initially derived via blocked cross-validation for the original sample and subsequently used for each block-bootstrapped sample. Unless stated differently, individual DFTs in

DFFs are post-pruned and the estimations of the parametric models in equation (2) and (3) build upon Stock and Watson (2012) and Siliverstovs and Wochner (2021) and use factor-augmented autoregressive processes with $L = 2$ lags and $R = 5$ factors.

2.4. Benchmarks

As dynamic factor trees and forests bridge dynamic factor models with regression trees and forests, we shall compare DFTs and DFFs against these two main benchmarks. Additionally, we also examine an extension of the DFM model as well as five conventional benchmarks typically employed in the relevant literature (e.g. H. H. Kim & Swanson, 2014; Siliverstovs & Wochner, 2021; Stock & Watson, 2012).

2.4.1. Main Benchmarks

Dynamic Factor Models (DFM)

Our first benchmark model of interest is the standard DFM,

$$Y_{t+h}^{(h)} = \boldsymbol{\psi}' \boldsymbol{W}_t^{(1)} + \boldsymbol{\phi}' \boldsymbol{F}_t^{(1)} + \epsilon_{t+h}^{(h)} \tag{4}$$

where parameter coefficients are estimated from the original sample (OS) and use $L = 2$ autoregressive lags as well as $R = 5$ factors, unless stated differently (e.g. Stock & Watson, 2002b, 2006, 2016; Siliverstovs & Wochner, 2021). In analogy to DFFs above, we will also consider a bootstrapped version, where we estimate a standard dynamic factor model for each bootstrapped sample (BS) (yielding $\boldsymbol{\psi}_b$ and $\boldsymbol{\phi}_b$) and subsequently average among all bootstrapped model coefficients. These models will be abbreviated as OS-DFM and BS-DFM, respectively.

Regression Trees (RT) and Random Forests (RF)

Conventional regression trees follow an algorithmic procedure that recursively optimizes a (least quares) objective function by repeatedly sub-dividing the predictor space spanned by the (typically large number of) variables in $\boldsymbol{V}_t^{(1)} \in \mathbb{R}^M$ into two mutually exclusive regions $\mathcal{R}$ and fitting a constant model to each of these until some stopping criterion is reached (e.g. the minimum quantity of observations per region η) and can be formally described as a piecewise constant model of the form,

$$\mu_{t+h}^{(h)} = \sum_{s=1}^{S} \varphi_s \mathbb{1}\big(\boldsymbol{V}_t^{(1)} \in \mathcal{R}_s(\eta)\big) \tag{5}$$

with $\mathcal{R}_s$ again denoting region $s \in \{1, \ldots, S\}$ (e.g. Breiman et al., 1984, James et al., 2013, p. 303ff. and Hastie et al., 2009, p. 307ff., for elaborate treatments; also Medeiros et al., 2021; Wochner, 2018). The best estimator for $\hat{\varphi}_s$ (in a least squares sense) corresponds to the sample average, $\hat{\varphi}_s = \frac{1}{T_s} \sum_{t \in \mathcal{R}_s} Y_{t+h}^{(h)}$, with $T_s = |\{t : \boldsymbol{V}_t^{(1)} \in \mathcal{R}_s\}|$ many observations in region $\mathcal{R}_s$ and $|\cdot|$ designating the size of the set (Hastie et al., 2009, p. 307f.; Breiman et al., 1984).

While the trees in forests are usually grown deep and have low bias albeit a high variance, random forests aim to reduce the increased variance not only by aggregating among numerous trees via bootstrap aggregation principles but also by decorrelating the trees via random predictor selections in each recursive splitting step (cf. Breiman, 1996a, 2001a; Hastie et al., 2009; James et al., 2013). The predictive equation of Breiman's (2001a) random forests can be expressed as,

$$\mu_{t+h}^{(h)} = \frac{1}{B} \sum_{b=1}^{B} \sum_{s=1}^{S} \varphi_{b,s} \mathbb{1}\big(\boldsymbol{V}_t^{(1)} \in \mathcal{R}_{b,s}(\eta)\big) \tag{6}$$

and thereby averages among B many (decorrelated) trees (cf. e.g. Hastie et al., 2009, p. 307ff., 587ff.). Similar to Medeiros et al. (2021), our empirical implementations of regression trees and random forests allow $\boldsymbol{V}_t^{(1)}$ to embrace a large set of M predictors; namely, all K predictors $\boldsymbol{X}_t^{(1)}$, the first $L = 2$ autoregressive terms in $\boldsymbol{W}_t^{(1)}$, as well as the first $R = 5$ factors in $\boldsymbol{F}_t^{(1)}$. Moreover, as regression trees are grown from the original sample and random forests from bootstrapped samples, we abbreviate the former as OS-RT and the latter as BS-RF. Specifically, with respect to BS-RFs, we follow analogously to the DFFs above the time series forest literature (e.g. Ehmann, 2020; Medeiros et al., 2021; Wochner, 2018) and employ (stationary) block-bootstrapped samples (cf. Canty, 2002; Politis & Romano, 1994).

2.4.2. Extended Benchmarks

Recession Probability Augmented Dynamic Factor Models (DFM-RP)

In an extension, we shall further consider another benchmark that is closely related to the regime-switching dynamic factor model estimated in Chauvet and Potter (2013),

$$Y_{t+h}^{(h)} = \boldsymbol{\psi}' \boldsymbol{W}_t^{(1)} + \boldsymbol{\phi}' \boldsymbol{F}_t^{(1)} + \boldsymbol{\vartheta}' \boldsymbol{Z}_t^{(1)} + \epsilon_{t+h}^{(h)} \tag{7}$$

where we directly include our partitioning variable as an explanatory variable into the DFM model (ibid., see their equation (16) on p. 164; also see references therein). While Chauvet and Potter (2013) use Markov-switching processes, we employ related albeit exogenously provided recession probability indices as partitioning variables (see Section 3 and 4). In analogy to the previous section, we will also consider a bootstrapped version, where we estimate equation (7) for each bootstrapped sample (yielding $\boldsymbol{\psi}_b$, $\boldsymbol{\phi}_b$, and $\boldsymbol{\vartheta}_b$) and then average among all bootstrapped model coefficients. We shall abbreviate the corresponding models as OS-DFM-RP and BS-DFM-RP, respectively.

2.4.3. Common Benchmarks

Similar to recent work (e.g. Medeiros et al., 2021; Siliverstovs, 2017b; Siliverstovs & Wochner, 2021; Stock & Watson, 2012; Wochner, 2018), we employ the following five common benchmarks: Historic mean (HMN), $Y_{t+h}^{(h)} = \varphi + \epsilon_{t+h}^{(h)}$, autoregressive processes with either a two, four or BIC-based lag order L (AR2, AR4,

ARL), $Y_{t+h}^{(h)} = \boldsymbol{\psi}' \boldsymbol{W}_t^{(1)} + \epsilon_{t+h}^{(h)}$, as well as combined autoregressive distributed lag models (CADL). The CADL model first estimates for each explanatory variable in $\boldsymbol{X}_t^{(1)}$ an autoregressive distributed lag model, $Y_{t+h}^{(h)} = \boldsymbol{\psi}' \boldsymbol{W}_t^{(1)} + \boldsymbol{\xi}' \boldsymbol{U}_{k,t}^{(1)} + \epsilon_{t+h}^{(h)}$ with $\boldsymbol{U}_{k,t}^{(1)} = (X_{k,t}, \ldots, X_{k,t-L+1})'$ and $L = 2$ for both autoregressive and distributed lag terms, and subsequently averages the predictions of these models among all $k \in \{1, \ldots, K\}$ (H. H. Kim & Swanson, 2014; Siliverstovs & Wochner, 2021).

2.5. Forecast Evaluation

We will assess forecasting performance in terms of relative Root Mean Squared Forecast Errors (RMSFE) of any two models m and b as follows,

$$\text{relative RMSFE}_{m,b}^{(h)} = \frac{\text{RMSFE}_m^{(h)}}{\text{RMSFE}_b^{(h)}}$$

where,

$$\text{RMSFE}_i^{(h)} = \left[\frac{1}{T - Q - h + 1} \sum_{t \in \mathscr{S}_f} \left(Y_{t+h}^{(h)} - \hat{Y}_{i,t+h}^{(h)} \right)^2 \right]^{1/2}$$

where $\hat{Y}_{i,t+h}^{(h)}$ denotes the h-period ahead prediction of model $i \in \{m, b\}$ (e.g. Korobilis, 2017; Hyndman & Koehler, 2006; Siliverstovs & Wochner, 2021; Stock & Watson, 2012). Moreover, we assess superior predictive ability of our dynamic factor trees and forests against the standard dynamic factor model by means of directed Diebold Mariano (1995) tests,

$$\text{H}_0: \ \text{E}\left[\left(\epsilon_{b,t+h}^{(h)} \right)^2 \right] = \text{E}\left[\left(\epsilon_{m,t+h}^{(h)} \right)^2 \right] \qquad \text{H}_A: \ \text{E}\left[\left(\epsilon_{b,t+h}^{(h)} \right)^2 \right] > \text{E}\left[\left(\epsilon_{m,t+h}^{(h)} \right)^2 \right]$$

and install, as in Siliverstovs and Wochner (2021), two cautionary measures through the use of heteroscedasticity and autocorrelation robust standard errors (for $h > 1$) as well as McCracken's (2007) critical values for nested model comparisons. Borrowing McCracken's (2007, p. 724) argument, directed testing is applied because our main models, m, (DFT, DFF) and main benchmark, b, (DFM) are nested.[5] Finally, building upon Welch and Goyal (2008), the state-dependent evaluation literature promotes the use of the cumulative sum of squared forecast error differences (CSSFED) between two models b and m,

$$\text{CSSFED}_{b,m}^{(h)}(t_0, t_1) = \sum_{t=t_0}^{t_1} \left(\hat{\epsilon}_{b,t+h}^{(h)} \right)^2 - \left(\hat{\epsilon}_{m,t+h}^{(h)} \right)^2$$

where $\hat{\epsilon}_{i,t+h}^{(h)} = Y_{t+h}^{(h)} - \hat{Y}_{i,t+h}^{(h)}$ and with continuously increasing t_0, which allows us to assess the forecast performance over time (see e.g. Siliverstovs, 2017a, 2020; Siliverstovs & Wochner, 2021). While a horizontal movement of CSSFED indicates similar performance between model m and b, an upward [downward] trending series indicates persistent superiority [inferiority] of model m over b, whereas an upward [downward] jumping series indicates transient superiority [inferiority] (e.g. Siliverstovs, 2017a; Siliverstovs & Wochner, 2021).

3. Data

We bridge three distinct FRED datasets for our main analyses all of which are provided by the Federal Reserve Bank of St. Louis. We use 125 monthly indicators from FRED-MD and the quarterly GDP time-series from FRED-QD as dependent variable, all of which are available from 1960 until 2018 (McCracken & Ng, 2019a, 2019b; see McCracken, 2019).[6] FRED-MD and FRED-QD constitute fairly new data services ever since 2015 and 2018, respectively, that manage data revisions and real-time updates (cf. McCracken & Ng, 2016, 2020; also McCracken, 2019) and find increasing use among macroeconomists (e.g. Korobilis, 2017; Medeiros et al., 2021; Siliverstovs & Wochner, 2021; Wochner, 2018). The third FRED source corresponds to the Econbrowser Recession Indicator (ERI)-Index, which builds upon the work of Marcelle Chauvet and James Hamilton (2006) and provides the probability of a recession at any given quarter since October 1967 (see FRED, 2019; Hamilton, 2018).[7] Hence, a balanced panel with information from all three sources is available as of October 1967.

Similar to Siliverstovs and Wochner (2021), the following data transformations were applied: First, less than 0.35% of the datasets were classified as outliers in FRED-MD and these were substituted with the median of the previous five observations (see Stock & Watson, 2012, online appendix B) and no outliers were detected for the GDP series in FRED-QD. Second, all data were stationarity transformed as defined in McCracken and Ng (2019a; 2019b) and all recession probability indices in use were first differenced. Following Stock and Watson (2012), the dependent variable was h-period stationarity transformed (ibid., see their Table B.2) as indicated by the super-index (h) in $Y_{t+h}^{(h)}$.

To cope with mixed frequencies, we follow Kim and Swanson (2014) who proposed to interpolate quarterly GDP values. As is well known, interpolations, however, may induce a measurement error and can affect the dynamics of the interpolated series and its associations with other variables (Angelini, Henry, & Marcellino, 2006, and references therein). To mitigate the concerns that a particular choice of interpolation method is driving our results, we will examine in total three distinct interpolation methods and will also assess the results in quarterly frequency as a robustness test (see Section 4.3). Our main specifications in Section 4.1 interpolate GDP (in levels) via Denton-Cholette (DCO) (Sax & Steiner, 2013 for a summary; Denton, 1971 and Dagum & Cholette, 2006 for detailed treatments).[8] Finally, similar to previous research (e.g. Siliverstovs & Wochner, 2021; Wochner, 2018), we assume the same publication structure for all variables as in the last vintage date and pursue a quasi real-time forecasting exercise[9] and for variables with ragged edges, we align the observations via Altissimo et al.'s (2010) lagging procedure.

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
			DFT-NUM(F)	**<u>0.773</u>*****	DFF-NUM(F)	**<u>0.741</u>*****
	AR4		DFT-BIN40(F)	**0.797*****	DFF-BIN40(F)	**0.772*****
	0.924		DFT-BIN50(F)	**0.795*****	DFF-BIN50(F)	**0.764*****
			DFT-BIN60(F)	**0.802*****	DFF-BIN60(F)	**0.767*****
	ARL	Targeted Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
			DFT-NUM(TF)	**<u>0.772</u>*****	DFF-NUM(TF)	**<u>0.719</u>*****
	CADL		DFT-BIN40(TF)	**0.789****	DFF-BIN40(TF)	**0.746*****
	0.922		DFT-BIN50(TF)	**0.779****	DFF-BIN50(TF)	**0.741*****
			DFT-BIN60(TF)	**0.787****	DFF-BIN60(TF)	**0.740*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
			DFT-NUM(F)	**0.830*****	DFF-NUM(F)	**<u>0.801</u>*****
	AR4		DFT-BIN40(F)	**0.853****	DFF-BIN40(F)	**0.824*****
	0.970		DFT-BIN50(F)	**<u>0.827</u>*****	DFF-BIN50(F)	**0.824*****
			DFT-BIN60(F)	**0.829*****	DFF-BIN60(F)	**0.824*****
	ARL	Targeted Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
			DFT-NUM(TF)	**0.845****	DFF-NUM(TF)	**<u>0.786</u>*****
	CADL		DFT-BIN40(TF)	**0.849****	DFF-BIN40(TF)	**0.820*****
	0.943		DFT-BIN50(TF)	**0.840*****	DFF-BIN50(TF)	**0.816*****
			DFT-BIN60(TF)	**<u>0.820</u>*****	DFF-BIN60(TF)	**0.813*****

Notes: The table entries show the relative root mean squared forecast error (RMSFE) of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-Index partitioning variable). DFT and DFF models are pruned via BIC criterion. The star symbols indicate the level of statistical significance from a one-sided Diebold Mariano (1995) test that assesses superiority in predictive performance of dynamic factor trees and forests (namely, DFT(F), DFT(TF), DFF(F) and DFF(TF)) or conventional regression trees and forests (namely, OS-RT(F), OS-RT(TF), BS-RF(F), BS-RF(TF)) against the corresponding (italicized) standard dynamic factor model (namely, OS-DFM(F), OS-DFM(TF), BS-DFM(F), and BS-DFM(TF)). For example, superior predictive ability of the DFT(TF) model is assessed against the OS-DFM(TF) model. The HMN, AR4, ARL and CADL benchmarks are compared against DFM(F). For nested model comparisons, McCracken's (2007) critical values are employed. The three FRED datasets outlined in Section 3 are used. The entries for DFTs and DFFs are bold if they have equal or lower MSFE than the corresponding DFM. The best model of each of the four groups per horizon is underlined. Factors are targeted based on a hard-threshold in the spirit of Bai and Ng (2008). The symbols *, **, *** indicate significance at the 10%, 5% and 1% level, respectively. For more details, see Section 4.1.

Table 1: Main Results (relative RMSFE)

4. Results

4.1. Main Results

The main results of our direct out-of-sample forecasting experiment for horizons $h = 1$ (nowcasts) and $h = 3$ (forecasts) is based on a recursively expanding scheme over a forecasting window of slightly more than 20 years (Jan. 1998 until Sep. 2018). Our proposed models can be divided into four main groups depending on the model type (DFT vs. DFF) and factor targeting (plain factors (F) vs. targeted factors (TF)). Within each of these four categories, we further distinguish between numeric (NUM) and binary measurement scales of the partitioning variable, where the latter discretize $\boldsymbol{Z}_t^{(1)}$ by assigning a binary classification to the values above and below the 40, 50, and 60th percentile of the partitioning variable, respectively (BIN40, BIN50, BIN60). Such a binary discretization (exogenously)

determines "good" and "bad" states, which limits the number of possible splits and may mitigate overfitting risks, particularly when the imposed regularizations appear to be weak (see Section 2 and 4.3; cf. Zeileis et al., 2008; also Strobl et al., 2011 and Zeileis & Hothorn, 2015).

4.1.1. Full Sample Evaluation

Table 1 summarizes our main results in terms of relative RMSFE against the AR2 benchmark over the full evaluation sample and highlights several interesting results: First, the dynamic factor tree and forest entries are printed in bold if they are superior in terms of RMSFE than the corresponding dynamic factor model. An inspection of Table 1 provides strong evidence in favor of the proposed models and highlights that dynamic factor trees and forests are statistically significantly superior to the standard dynamic factor model. In terms of MSFE, dynamic factor trees and forests tend to improve upon the standard dynamic factor models by over 20%. For example, the best performing dynamic factor forest for three-month ahead predictions (DFF-NUM(TF)) achieves a relative MSFE of 0.617 whereas the standard targeted dynamic factor model from bootstrapped sample (BS-DFM(TF)) has a relative MSFE of 0.796. Second, the results also appear to be well in accord with the targeting literature (e.g. Bai & Ng, 2008; Stock & Watson, 2012; Boivin & Ng, 2006) in that targeting tends to improve predictive performance. Moreover, numeric rather than binary partitions are often able to meaningfully exploit the richer information and perform slightly superior. Third, consistent with the tree-based ensemble and forecast combination literature (e.g. Medeiros et al., 2021; Breiman, 1996a, 2001a; James et al., 2013, Hastie et al., 2009 and Elliott & Timmermann, 2016, for reviews), dynamic factor forests systematically improve upon dynamic factor trees in all cases. Fourth, we notice that DFTs and DFFs can systematically improve upon standard regression trees and random forests. As is also shown, regression trees perform rather poorly, which was to be expected because of the high risks of overfitting whereas random forests successfully overcome these concerns (cf. Breiman, 1996a, 2001a; e.g. James et al., 2013, for a discussion).

4.1.2. Sub-Sample Evaluation

To better understand the evolution of these forecasting improvements, we follow the burgeoning state-dependent forecast evaluation literature and examine (in addition to full sample forecast evaluations above) also those for the subsamples in boom and bust periods according to NBER (2020) (e.g. Chauvet & Potter, 2013; Fossati, 2018; Siliverstovs, 2017a; Siliverstovs & Wochner, 2021). Table 2 reveals that the DFT and DFF models perform better than the DFM in both recessionary and expansionary subsamples. Specifically, during recessions they typically achieve seizable and statistically significant improvements over the DFM, whereas in expansions the improvements are still present but more moderate in terms of size and significance. Especially at higher forecasting horizons, the performance during expansions is similar to those of the AR2 benchmark,

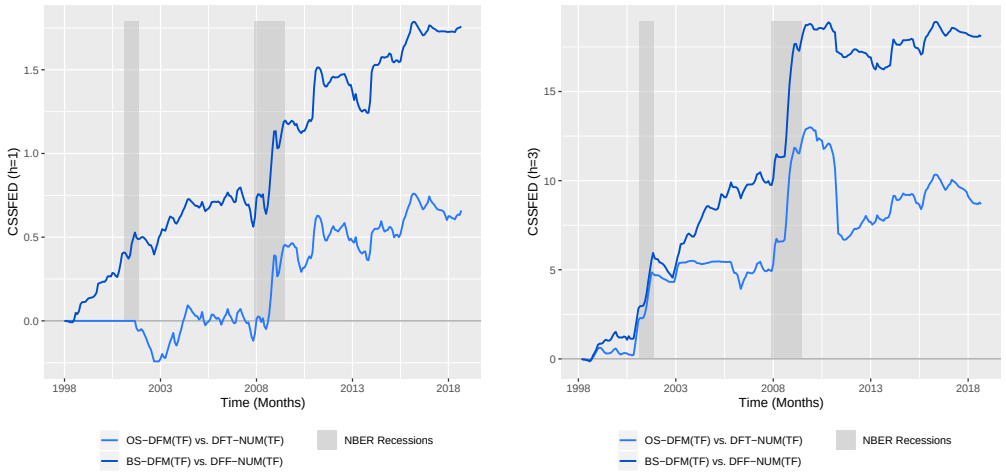
Horizon		BM Results		Dynamic Factor Trees Models	rRMSFE	Dynamic Factor Forests Models	rRMSFE
h=1	Expansion	HMN	Factors	OS-DFM(F)	*0.979*	BS-DFM(F)	*0.976*
		0.991		OS-RT(F)	1.451	BS-RF(F)	0.948
		AR4		DFT-NUM(F)	**0.939*****	DFF-NUM(F)	**0.888*****
		0.966		DFT-BIN50(F)	**0.968**	DFF-BIN50(F)	**0.921*****
		ARL	Trgt. Fact.	OS-DFM(TF)	*0.941*	BS-DFM(TF)	*0.939*
		0.936+		OS-RT(TF)	1.584	BS-RF(TF)	0.950
		CADL		DFT-NUM(TF)	**0.930**	DFF-NUM(TF)	**0.861*****
		0.956		DFT-BIN50(TF)	0.943	DFF-BIN50(TF)	**0.887*****
	Recession	HMN	Factors	OS-DFM(F)	*0.655*	BS-DFM(F)	*0.654*
		1.012		OS-RT(F)	0.872	BS-RF(F)	0.744
		AR4		DFT-NUM(F)	**0.533*****	DFF-NUM(F)	**0.535*****
		0.877		DFT-BIN50(F)	**0.544*****	DFF-BIN50(F)	**0.539*****
		ARL	Trgt. Fact.	OS-DFM(TF)	*0.613*	BS-DFM(TF)	*0.612*
		0.847		OS-RT(TF)	0.962	BS-RF(TF)	0.725
		CADL		DFT-NUM(TF)	**0.549****	DFF-NUM(TF)	**0.520*****
		0.883		DFT-BIN50(TF)	**0.544*****	DFF-BIN50(TF)	**0.534*****
h=3	Expansion	HMN	Factors	OS-DFM(F)	*1.067*	BS-DFM(F)	*1.064*
		1.003**		OS-RT(F)	1.819	BS-RF(F)	1.023
		AR4		DFT-NUM(F)	**1.028****	DFF-NUM(F)	**1.003*****
		0.999		DFT-BIN50(F)	**1.028****	DFF-BIN50(F)	**1.025****
		ARL	Trgt. Fact.	OS-DFM(TF)	*1.058*	BS-DFM(TF)	*1.059*
		1.000		OS-RT(TF)	1.740	BS-RF(TF)	1.013
		CADL		DFT-NUM(TF)	1.068	DFF-NUM(TF)	**0.982*****
		0.971+		DFT-BIN50(TF)	**1.043**	DFF-BIN50(TF)	**1.005****
	Recession	HMN	Factors	OS-DFM(F)	*0.769*	BS-DFM(F)	*0.761*
		1.023		OS-RT(F)	1.116	BS-RF(F)	0.894
		AR4		DFT-NUM(F)	**0.623*****	DFF-NUM(F)	**0.587*****
		0.946		DFT-BIN50(F)	**0.616*****	DFF-BIN50(F)	**0.612*****
		ARL	Trgt. Fact.	OS-DFM(TF)	*0.735*	BS-DFM(TF)	*0.730*
		0.954		OS-RT(TF)	1.044	BS-RF(TF)	0.898
		CADL		DFT-NUM(TF)	**0.605*****	DFF-NUM(TF)	**0.576*****
		0.919		DFT-BIN50(TF)	**0.629*****	DFF-BIN50(TF)	**0.619*****

Notes: Building upon Chauvet and Potter (2013) and Siliverstovs and Wochner (2021), the table entries show the relative RMSFE of a particular model against the AR2 benchmark for expansionary and recessionary sub-samples separately (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-Index partitioning variable). Recessions and expansions are determined according to NBER (2020). For $h = 1$ and $h = 3$, the forecast evaluation window contains 249 and 247 monthly time periods, respectively; the expansionary sub-sample contains 221 and 219 observations and the remaining observations belong to the recessionary sub-sample. BM stands for benchmarks. For more details about table entries, see notes in Table 1.

Table 2: Main Results for Sub-Samples (relative RMSFE)

which is again in accord with the state-dependent forecast evaluation literature (cf. e.g. Siliverstovs & Wochner, 2021).

Building upon Siliverstovs (2017a; 2020) and Siliverstovs and Wochner (2021), Figure 1 provides the CSSFED of the dynamic factor model from original samples (b_1 = OS-DFM(TF)) against the targeted dynamic factor tree (m_1 = DFT-NUM(TF)) as well as the CSSFED of the dynamic factor model from bootstrapped samples (b_2 = BS-DFM(TF)) against the targeted dynamic factor forest (m_2 =



1.a: Nowcast ($h = 1$) **1.b:** Forecast ($h = 3$)

Notes: This figure shows the CSSFED of the conventional dynamic factor model against the top-performing dynamic factor trees and forests in Table 1 and displays recession classifications of NBER (2020) (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). An upward {horizontal} [downward] movement of CSSFED indicates superior {equal} [inferior] performance of DFT or DFF over DFM (Siliverstovs, 2017a, 2020; Siliverstovs & Wochner, 2021). While the absolute levels of two CSSFED in Figure 1.a and 1.b may not be compared quantitatively because the CSSFED are not measured on similar scales (non-standardized SFED) (cf. Siliverstovs & Wochner, 2021), they reveal qualitatively similar dynamics. The figure is based on the three FRED datasets (see Section 3).

Figure 1: DFTs and DFFs Relative Performance against DFMs over Time (CSSFED)

DFF-NUM(TF)). As explained in Section 2.5, the figure visualizes how the relative performance of these models evolves over the evaluation window: The two series show an upward trending behavior as well as larger upward jumps during recessions (with slight to pronounced deterioriations in the aftermath of recessions). At the same time, we also see considerable differences between the CSSFED series for DFTs and DFFs and when and how DFFs are able to improve upon DFTs. In short, while both DFTs and DFFs can improve upon DFMs, the superiority in DFFs performance is not only more pronounced but also more persistent than the one of DFTs.

4.2. Extended Results

We examine additionally the following extensions: First, we consider our augmented DFM benchmarks (see Section 2.4.2). Second, we extend the previous analyses by keeping the same parametric form in our DFTs and DFFs (see Section 2.3) but allow the partitioning sets to take more complex multi-variate forms so as to allow for generalized and time-varying state-dependent dynamics. Specifically, we consider, on the one hand, in analogy to the structural change point literature (cf. Goulet Coulombe, 2020, Zeileis et al., 2008, and references therein; also Audrino & Medeiros, 2011) the inclusion of time in addition to the ERI-Index as partitioning variable and label the corresponding models DFT-NUM-PLUS-T and DFF-NUM-PLUS-T, respectively. On the other, as the forest literature often considers a large number of predictors (e.g. Goulet Coulombe, 2020; Medeiros et

al., 2021), we incorporate in addition to the ERI-Index all contemporaneous explanatory variables and refer to the corresponding models as DFT-NUM-PLUS-X and DFF-NUM-PLUS-X, respectively. Third, we examine the impact of using Lasso-based penalization to enable differing state-dependent model designs (cf. Quan et al., 2020). Finally, we examine the use of novel factors within the proposed framework of DFFs (see Section 2.1).

4.2.1. Recession-Probability Augmented DFM

The results of the first two extensions are shown in Table 3 together with some results from Table 1 for better comparability. First, we notice that dynamic factor forests (DFF-NUM) perform about equally well as the augmented dynamic factor benchmarks (DFM-RP) — which is broadly in line with Chauvet and Potter (2013) in that DFM-RP models are closely related to their Markov-switching dynamic factor model, which also achieved strong predictive performance in their setting.

4.2.2. Generalized and Time-Varying State-Dependent Dynamics

Second, we find that DFTs and DFFs with multivariate partitioning sets are generally able to outperform the corresponding DFMs with DFFs showing again more robust performance than DFTs. However, relative to the univariate settings in Table 1, larger partitioning sets can be both helpful and harmful for performance: While the augmentation with T can lead to notable additional improvements of DFF-NUM-PLUS-T over DFF-NUM (especially at $h = 3$), the augmentation with X generally cause the performance of DFF-NUM-PLUS-X to drop below that of DFF-NUM. Nonetheless, even in the latter case, DFF-NUM-PLUS-X is still better than the DFM. These results appear generally comparable to Goulet Coulombe (2020) as well as broadly consistent with the guiding principle in high-dimensional settings that the addition of relevant features is generally beneficial whereas the addition of irrelevant features can have unfavourable consequences for out-of-sample performance (see discussion in James et al., 2013, p. 241ff.).

4.2.3. Ridge- and Lasso-based Penalization

Third, Table 4 shows the results for the alternative regularization scheme inspired by Quan et al. (2020): As can be seen, the penalized DFTs and DFFs can indeed yield more accurate results than their non-penalized and unpruned counterparts. In other words, the regularization appears to bear positive effects. Moreover, when comparing the performance of this kind of regularization with the post-pruning results in Table 1, we realize that Ridge- and Lasso-based penalization may even perform superior to BIC-based tree pruning in some (but not all) cases.

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-DFM-RP(F)	0.773***	BS-DFM-RP(F)	0.772***
			DFT-NUM(F)	**0.773*****	DFF-NUM(F)	**0.741*****
	AR4		DFT-NUM-PLUS-T(F)	**0.770*****	DFF-NUM-PLUS-T(F)	**0.740*****
	0.924		DFT-NUM-PLUS-X(F)	**0.834***	DFF-NUM-PLUS-X(F)	**0.788*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-DFM-RP(TF)	0.755***	BS-DFM-RP(TF)	0.754***
			DFT-NUM(TF)	**0.772*****	DFF-NUM(TF)	**0.719*****
	CADL		DFT-NUM-PLUS-T(TF)	**0.779****	DFF-NUM-PLUS-T(TF)	**0.714*****
	0.922		DFT-NUM-PLUS-X(TF)	**0.784*****	DFF-NUM-PLUS-X(TF)	**0.757*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-DFM-RP(F)	0.781***	BS-DFM-RP(F)	0.780***
			DFT-NUM(F)	**0.830*****	DFF-NUM(F)	**0.801*****
	AR4		DFT-NUM-PLUS-T(F)	**0.832****	DFF-NUM-PLUS-T(F)	**0.788*****
	0.970		DFT-NUM-PLUS-X(F)	**0.853*****	DFF-NUM-PLUS-X(F)	**0.855*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-DFM-RP(TF)	0.775***	BS-DFM-RP(TF)	0.774***
			DFT-NUM(TF)	**0.845****	DFF-NUM(TF)	**0.786*****
	CADL		DFT-NUM-PLUS-T(TF)	**0.829****	DFF-NUM-PLUS-T(TF)	**0.757*****
	0.943		DFT-NUM-PLUS-X(TF)	0.930	DFF-NUM-PLUS-X(TF)	**0.831*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The acronyms PLUS-T and PLUS-X refer to sets of splitting variables that incorporate, in addition to the ERI-Index, time (T) and all contemporaneous explanatory variables (X), respectively. For more details, see notes in Table 1.

Table 3: Extended Results: Generalized Partitioning Sets (relative RMSFE)

4.2.4. Generalized and Time-Varying State-Dependent Dynamics with Lasso-based Penalization

Fourth, Table A.1 combines the two previous extensions and examines both Lasso-based regularization and generalized partitioning sets. While the results remain overall fairly comparable to the ones above, we notice that BIC-based pruning appears superior to Lasso-based pruning when too many splitting variables (PLUS-X) are available. Interestingly, however, we notice that the Lasso in combination with time-augmented partitioning sets (PLUS-T) typically achieves the very best results among the ones considered so far. Specifically, the time-augmented DFFs in Table A.1 with Lasso-based regularization systematically outperform the corresponding DFFs of the main results in Table 1 (neither time-augmentation nor Lasso-based penalization), the ones provided in the extended results in Table 3 (time-augmented partitioning set) as well as those of Table 4 (Lasso-based penalization). This underpins the importance of allowing both generalized time-varying state-dependent dynamics as well as regime-switching behaviour between structurally distinct models.

	Benchmarks		Dynamic Factor Trees		Dynamic Factor Forests	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
			DFT-NUM(F)	**0.772*****	DFF-NUM(F)	**0.753*****
	AR4		DFT-NUM-RIDGE(F)	**0.769*****	DFF-NUM-RIDGE(F)	**0.751*****
	0.924		DFT-NUM-LASSO(F)	**0.771*****	DFF-NUM-LASSO(F)	**0.750*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
			DFT-NUM(TF)	**0.755*****	DFF-NUM(TF)	**0.730*****
	CADL		DFT-NUM-RIDGE(TF)	**0.751*****	DFF-NUM-RIDGE(TF)	**0.733*****
	0.922		DFT-NUM-LASSO(TF)	**0.748*****	DFF-NUM-LASSO(TF)	**0.729*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
			DFT-NUM(F)	**0.844*****	DFF-NUM(F)	**0.816*****
	AR4		DFT-NUM-RIDGE(F)	**0.842*****	DFF-NUM-RIDGE(F)	**0.816*****
	0.970		DFT-NUM-LASSO(F)	**0.835*****	DFF-NUM-LASSO(F)	**0.813*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
			DFT-NUM(TF)	**0.853****	DFF-NUM(TF)	**0.803*****
	CADL		DFT-NUM-RIDGE(TF)	**0.850****	DFF-NUM-RIDGE(TF)	**0.807*****
	0.943		DFT-NUM-LASSO(TF)	**0.839*****	DFF-NUM-LASSO(TF)	**0.800*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; no pruning). The DFTs and DFFs do neither apply pre- nor post-pruning but apply Ridge and Lasso regularization. For more details, see notes in Table 1.

Table 4: Extended Results: Penalized DFTs and DFFs (relative RMSFE)

4.2.5. Novel Factors

Table A.2 examines the impact of employing novel rather than traditional factors in our state-dependent models (see Section 2.1). While we find that all three tend to perform generally well in that the DFTs and DFFs appear to be generally superior to DFMs, we do not find any material improvements over PCA-based factors. In fact, we find PLS- and SPCA-based factors to perform generally somewhat worse whereas KPCA-based factors can perform slightly better than PCA-based factors. Interestingly, we find that targeting tends to be generally helpful for PLS-, KPCA- as well as SPCA-based factors. Table A.3 takes the novel factor examination one step further and assesses the time-augemented partitioning set together with Lasso-based penalization. Relative to the results shown in Table A.2, PLS shows markedly better performance in all cases and DFFs also show superior results, which indicates that time-augmented partitioning sets together with Lasso-based penalization also appear to be useful for novel factors.

4.2.6. Two-Stage Machine Learning Framework

One of the key ingredients to the previous analyses was the recession probability indicator (ERI-Index), which we retrieved from external sources (see Section 3 for details). In a sense, the approach outlined until this point may be considered as

a single-stage machine learning approach. In a final key extension, we build upon recent work and extend our approach to a two-stage machine learning framework by internalizing the prediction of the recession probability in a first stage before using it within DFTs and DFFs in a second stage.

Our first stage builds directly upon the recent work in Ng (2014), Döpke et al. (2017) and Pierdzioch and Gupta (2020) (among others) who proposed to use machine learners such as boosted regression trees to predict the probability of a recession in data-rich environments. We will consider in addition to Boosted Regression Trees (BRT), Random Forests (RF) as another non-linear classifier as well as Lasso (LAS) as a regularized linear classification technique to produce predicted recession probabilities (PRP) (see e.g. Friedman, 2001, 2002, for boosting, Breiman, 2001a, for random forests and Tibshirani, 1996, for lasso; see Hastie et al., 2009 for an excellent textbook treatment of all three methods). All three methods are trained to classify monthly NBER (2021) recessions from the set of contemporaneous explanatory variables and are tuned via cross-validation to yield the best receiver operating characteristic (ROC) curves, which are robust to changing class distributions (Fawcett, 2006; see Kuhn, 2008, 2019, for implementation details). As discussed in Ng (2014), however, machine learning procedures such as boosting may yield insufficiently persistent recession probability predictions with only little or no autoregressive dynamics at all (also see Chauvet & Hamilton, 2006 and Hamilton, 2018 for distinct yet related discussions about smoothened PRPs). An arguably simple yet potentially expedient approach to deal with this issue at least partially is a subsequent smoothing step that applies trend filtering techniques. Specifically, we use the Hodrick-Prescott (1997) (HP) filter to generate a smoother series of predicted recession probabilities (see Balcilar, 2019 for implementation details). Our second stage then subsequently uses the smoothed PRP within DFTs and DFFs[10], possibly in combination with generalized partitioning sets and state-dependent regularization.

Table 5 shows the result from using our internalized PRP instead of the ERI-Index. As is evident from the table, DFTs and DFFs tend to yield more accurate predictions than our main autoregressive benchmark in all cases and may produce even significantly better predictions than their DFM counterparts, especially for one-step ahead predictions. Among the different PRP methods used, random forests appear to have a slight advantage. Table A.15 shows the results for our internalized PRP in combination with a generalized partitioning set that additionally includes time as well as Lasso-based state-dependent penalizations. At least for one-step ahead predictions, DFFs still tend to produce substantially and significantly better results than DFMs but under these specifications no longer for multi-step ahead predictions. Moreover, in direct comparison with the previous results for the ERI-Index, we find that the ERI-Index appears to perform generally even better than PRP-Indices. Nonetheless, we perceive the results for PRP to be certainly encouraging as they show that the extension to a two-stage machine learning framework can also improve upon DFMs in statistically significant ways.

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
			DFT-NUM-PRP[BRT](F)	**0.838**	DFF-NUM-PRP[BRT](F)	**0.825****
	AR4		DFT-NUM-PRP[RF](F)	**0.806*****	DFF-NUM-PRP[RF](F)	**0.792*****
	0.924		DFT-NUM-PRP[LAS](F)	**0.827***	DFF-NUM-PRP[LAS](F)	**0.792*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
			DFT-NUM-PRP[BRT](TF)	**0.789*****	DFF-NUM-PRP[BRT](TF)	**0.782****
	CADL		DFT-NUM-PRP[RF](TF)	**0.770*****	DFF-NUM-PRP[RF](TF)	**0.764*****
	0.922		DFT-NUM-PRP[LAS](TF)	0.820	DFF-NUM-PRP[LAS](TF)	**0.777****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
			DFT-NUM-PRP[BRT](F)	0.948	DFF-NUM-PRP[BRT](F)	0.916
	AR4		DFT-NUM-PRP[RF](F)	0.964	DFF-NUM-PRP[RF](F)	**0.909**
	0.970		DFT-NUM-PRP[LAS](F)	**0.897***	DFF-NUM-PRP[LAS](F)	**0.901***
	ARL	Trgt. Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
			DFT-NUM-PRP[BRT](TF)	0.905	DFF-NUM-PRP[BRT](TF)	0.901
	CADL		DFT-NUM-PRP[RF](TF)	**0.879****	DFF-NUM-PRP[RF](TF)	**0.890***
	0.943		DFT-NUM-PRP[LAS](TF)	0.916	DFF-NUM-PRP[LAS](TF)	0.910

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; PRP-IDX partitioning variable; BIC-based pruning; DFFs use 200 bootstrapped samples). The acronym PRP indicates that the partitioning variable is based on predicted recession probabilities (two-stage machine learning framework), which are themselves derived either via boosted regression trees [BRT], random forest [RF] and lasso [LAS] algorithms and subsequently smoothened. For more details, see notes in Table 1.

Table 5: Extended Results - Two-Stage Machine Learning Framework: PRP-based DFTs and DFFs (relative RMSFE)

4.3. Robustness Results

This section examines how sensitive our estimation results are with respect to several modeling choices and assumptions. We will hereafter assess alternatives typically encountered in the relevant literature (e.g. H. H. Kim & Swanson, 2014; Siliverstovs & Wochner, 2021; Stock & Watson, 2012). We will incorporate the extended analyses with augmented partitions and penalized regressions for the robustness tests on the two alternative interpolation methods, rolling window estimations as well as extended forecasting windows.

4.3.1. Alternative Interpolations

To assess the robustness of our results with regards to the choice of a particular interpolation method, we shall examine three distinct ones: Our main specifications interpolate GDP (in levels) via Denton-Cholette (DCO), which seeks to preserve the movement at a higher frequency (cf. Sax & Steiner, 2013, and references therein; also Dagum & Cholette, 2006). Kim and Swanson (2014) propose to interpolate GDP via Chow-Lin (1971), which can either be applied to stationary or co-integrated series (Sax & Steiner, 2013, p. 80ff.). In the former case, GDP (in growth rates) is interpolated here from a dynamic factor model with monthly factors (CLU).[11] In the latter, GDP (in levels) is interpolated from three monthly co-integrated series (CL3).[12] All three interpolations (DCO, CL3, CLU) are implemented via Sax et al. (2020) and qualify as equally appropriate as there has not yet emerged a general consensus on a first best interpolation method (Guérin & Marcellino, 2013). Table A.4 shows the result for the CLU-based interpolations of GDP growth. As can be seen the results are fairly similar to DCO-interpolations both in terms of size and significance (except for DFT(TF)s at $h = 1$ because the algorithm prunes the DFTs to a single-state model). Still DFFs outperform the DFMs and tend to be substantially better than conventional benchmarks (e.g. HMN, ARL). An analogous result holds for CL3 interpolations: Table A.5 shows that the main results can be qualitatively maintained but are quantitatively less pronounced in terms of size and significance than CLU and DCO. Interestingly, we also find that Lasso-based penalization can yield the best-performing DFT and DFF models for CLU at $h = 1$ and for CL3 at $h = 3$ and we notice that these improvements can be quite pronounced. Moreover, we find that the time-augmented partitioning sets tend to improve the predictive performance of DFTs for CLU at $h = 1$ and DFTs and DFFs for CL3 at $h = 3$.

4.3.2. Rolling Windows

While our main results are based on recursively expanding windows, Table A.6 examines the effect of using rolling schemes with window lengths of 300 months (see e.g. H. H. Kim & Swanson, 2014; Stock & Watson, 2012). The table shows that our main findings are robust to this change. In fact, the results strengthen our main findings in that the improvements of DFTs and DFFs over DFMs tend to be slightly more accentuated under rolling than under recursive windows.

Moreover, as opposed to our previous setting, the considered extensions of Lasso- and Ridge-based penalization as well as extended partitioning sets tend to work generally less well than the previous robustness test.

4.3.3. Extended Forecasting Window

While the forecast evaluation window for the main specification in Section 4.1 includes over 20 years (1998-2018) with almost 250 monthly predictions, we examine an extension of the evaluation window to 1985-2018 with over 400 monthly observations (cf. e.g. Siliverstovs & Wochner, 2021; Wochner, 2018). Table A.7 shows that the main results still persist but are weaker in terms of size (especially for nowcasts), which indicates that the nowcasting over the post-millennial period appears to be better than over the pre-millennial period (also see Wochner, 2018, for related findings). As the sources of these differences cannot be attributed to insufficient power due to shorter sample sizes (because DFFs tend to split over the pre-millennial era), they appear to be, at least in part, structural in the sense that the proposed methodology appears to work worse over the pre-millennial period (1985-2000).

Additionally, we notice that both Lasso and Ridge can yield some improvements for DFTs and DFFs generally benefit from having a time-augmented partitioning variable. The augmentation with a vast number of contemporary indicators performs generally still better than the standard DFM but generally worse than the more narrowly specified partitioning sets.

4.3.4. Quarterly Frequency

Table A.8 provides the estimation results for quarterly series by quarterly aggregation of the monthly FRED-MD series (cf. e.g. Foroni & Marcellino, 2013). These estimation results employ cross-validated minimum state sizes because we made the experience that information criteria can often regularize the models too heavily, forcing them frequently into a single state model, which leaves only little room for performance differences over and above the DFM (see discussions in Section 2). The results in Table A.8 paint an encouraging picture and indicate that our main findings extend fairly well to the quarterly frequency case (in case of CV-based regularization).

4.3.5. Pre-Pruning

As stated in Section 2, we examine both pre- and post-pruning strategies. In contrast to the post-pruning approach of our main results, Table A.9 highlights the results from a pre-pruning strategy. The two approaches appear to perform approximately equivalently well and leave our main findings quantitatively and qualitatively unaffected.

4.3.6. Single Factor and Ten Factors

While the main results are all based on five dynamic factors, Table A.10 and A.11 examine the model performance with only a single factor and ten factors (cf. Siliverstovs & Wochner, 2021; Stock & Watson, 2016). The ten factor model tends to perform slightly superior than the one factor model in terms of relative RMSFE over the AR2 benchmark. While the one factor model is able to split DFTs more frequently, the performance of DFFs appears to be quite comparable to the five factor model.

4.3.7. Alternative Recession Probability Indices

While our main specifications make use of the ERI-index (see Chauvet & Hamilton, 2006; Hamilton, 2018), the Federal Reserve Bank of Philadelphia (FRBP) conducts the Survey of Professional Forecasters (SPF) and provides the mean survey response concerning the probability of a recession in the current (SPF-REC1-IDX) and next quarter (SPF-REC2-IDX) ever since 1968 (see FRBP, 2018, 2019). These survey-based indicators shall be used as alternative partitioning variables to the data-based version of the ERI-index. In analogy to our main analyses, the quarterly SPF-REC1 and SPF-REC2 indices are interpolated to monthly frequencies via Denton-Cholette (cf. Sax & Steiner, 2013, and references therein). The results in Table A.12 and A.13 provide empirical support for the proposed dynamic factor trees and forests for these alternative recession probability indices: At all horizons the results are only slightly weaker and still outperform the standard DFM and rank in the majority of cases as the first best model among all those considered. This suggests that the proposed methodology is not dependent on the ERI-index and generalizes to alternative recession probability indicators.

4.3.8. Publication Lag

The ERI-Index has a fairly consistent publication structure and is released with a publication lag of 4 to 6 months.[13] While the main results assumed the ERI-index to be released together with all monthly indicators, i.e. having a one month lag, we may alternatively seek to predict the missing values arising from lagged publications (cf. e.g. Bulligan, Marcellino, & Venditti, 2015). Inspired by Kim and Swanson (2016), we infer the missing values on the current edge, $Z^{(1)}_{t+h}$, with a simple regime-switching model that alternates between predictions from a distributed lag model and a beta regression model (cf. Cribari-Neto & Zeileis, 2010; Hill, Griffiths, Lim, & Lim, 2011).[14] Table A.14 provides the estimation results and indicates that the timely release of the ERI-index contains indeed relevant information — particularly for the 3-months ahead forecasts. While the nowcasting results ($h = 1$) remain fairly robust to publication lags, the forecasting results ($h = 3$) are weaker.

These findings suggest that the employed ERI-index entails both advantages and disadvantages: On the one hand, it appears to be the one that yields the best performance in our modeling framework among all indices considered. On the

other, the index is unfortunately not (yet) available in a monthly frequency and publication lags appear to weaken model performance to some extent. In light of these results, a high-frequency real-time release of the ERI-index à la Aruoba, Diebold and Scotti's (2009) business cycle indicator appears to be desirable. In any case, the analyses have made clear that a weaker performance due to publication lags can only hardly be considered as a weakness of the proposed dynamic factor trees and forests but constitute rather a limitation of the data.

5. Conclusion

This study proposed dynamic factor trees and forests based on Zeileis et al.'s (2008) model-based recursive partitioning algorithm and its extension in the spirit of Garge et al. (2013), which allow to embed theory-led dynamic factor models within tree-based machine learning ensembles conditional on the state of the business cycle (see Section 1 and 2). Building upon recent advances outlined above, we also studied more complex conditioning sets to account for generalized and time-varying state dependent dynamics under forest-based randomization schemes, examined the inclusion of Lasso-based regularization to enable switching mechanisms between structurally distinct state-dependent models, studied novel factor extraction methods within the proposed DFF framework and extended the framework to two stages by internalizing the predictions of recession probabilities based on machine learning classifiers that may cope with data-rich environments. In our out-of-sample forecasting exercise for short-term GDP growth predictions, we generally find significant empirical evidence that DFFs can indeed systematically outperform standard dynamic factor models in both expansive and contractive periods. The results also qualify as fairly robust against a large number of robustness tests. These observations corroborate the idea that the proposed state-dependent models are indeed able to beneficially exploit the time-varying dynamics in good and bad times — much like sailors who benefit from maneuvering their ships differently in stormy and calm seas (cf. K. Kim & Swanson, 2016; Diebold & Rudebusch, 1996).

We see several possible avenues for future research: The generalizability of the proposed modelling framework will have to be examined for alternative dependent variables and countries. This is relevant because it is not certain à-priori that it will generalize as well to alternative settings because other series and states may exhibit structurally distinct characteristics: For instance, Corradi and Swanson's (2014) test for structural instability in dynamic factor models rejects the null of stability for US GDP growth but fails to reject the null for the S&P500, producer price index or the 10-year Treasury-bond rate. Likewise, Carstensen et al. (2020, p. 830f.) argue that cross-country differences in the persistence of the very same series can greatly affect the success of factor-based autoregressive processes across different countries. Additionally, we find the provision of a vast set of partitioning variables to be generally inferior than a carefully selected one and more research is needed to better guide its composition (also see Goulet Coulombe, 2020). Hence, future work along these lines will contribute towards a better comprehension of

the circumstances and environmental conditions needed to bring novel modeling architectures to fruition and may ultimately provide policy makers with better guidance in selecting the most appropriate tools for their decisions.

Notes

1. Theoretical equilibrium models often assume that macroeconomic dynamics are driven by a few shocks, such as technology or monetary policy shocks (Bai & Ng, 2007; Giannone, Reichlin, & Sala, 2006). Dynamic factor models are consistent with these assumptions and rest on the idea that the co-movements among many economic variables can be decomposed into two (orthogonal) sources: A "common component" (common shocks or common factors) that captures a few unobserved factors that govern the dynamics of many variables plus a "idiosyncratic component", which captures residual peculiarities of each individual series (e.g. D'Agostino, Giannone, Lenza, & Modugno, 2016; Diebold & Rudebusch, 1996; Forni et al., 2009; Giannone et al., 2006; Stock & Watson, 2002b, 2006, 2016, 2017). Furthermore, theoretical business cycle equilibrium models can be shown to take a factor-like structure in case of measurement errors (cf. e.g. Giannone et al., 2006, and Diebold & Rudebusch, 1996, and references therein).

2. Targeting for these factors is done analogously to the PCA-based case.

3. Equation (2) is, for instance, closely related to standard regime switching models, such as threshold autoregressive (TAR) models: Specifically, if there were no factors present and if the partitioning variable corresponded to lagged values of the dependent variable, then equation (2) would take a similar structure as the self-exciting threshold autoregressive model (cf. e.g. Kock & Teräsvirta, 2011, p. 62ff., and references therein).

4. Notice that the default minsize η in each terminal node is set to be ten times the number of estimated parameters, which also provides some cushion against overfitting (see Hothorn & Zeileis, 2020).

5. The determination of McCracken's (2007) critical values for two nested models requires the difference in the number of model parameters, k_2 (excess parameters). For λ_t many terminal nodes in the dynamic factor tree in period t, there are $(\lambda_t - 1) \times (1 + L + R)$ many excess parameters relative to the dynamic factor model (i.e. one additional intercept, L additional autoregressive lags and R additional factors per additional terminal node). To determine k_2 for trees [and forests], we use the average number of terminal nodes, $\bar{\lambda} = 1/|\mathscr{S}_f| \sum_{t=1}^{T} \lambda_t$ [and $\bar{\lambda} = 1/|\mathscr{S}_f| \sum_{t=1}^{T} 1/B \sum_{b=1}^{B} \lambda_{b,t}$]. Moreover, since the maximum k_2 provided by McCracken (2007) is limited to 10, we use $k_2 = \min(\bar{\lambda}, 10)$.

6. The raw datasets for FRED-MD actually provide information for 128 monthly and 248 quarterly macroeconomic indicators (see McCracken & Ng, 2019a, 2019b). From FRED-MD, we dropped monthly and quarterly datasets which are not available at the beginning of 1960 (cf. e.g. Wochner, 2018) and retained quarterly GDP from FRED-QD. Plain factors are subsequently extracted from the 375 contemporaneous, first and second order lags of the monthly indicators (cf. Ehmann, 2020; H. H. Kim & Swanson, 2014). Moreover, to retain the `UMCSENTx` index (FRED-MD), which is only available in quarterly frequency until Q4-1977, the initial values were interpolated via Denton-Cholette (Sax & Steiner, 2013, and references therein).

7. The FRED mnemonic of the index is `JHGDPBRINDX`. We will cope with the mixed-frequency problem by interpolating quarterly recession probability indices via Denton-Cholette (Sax & Steiner, 2013, and references therein) and will examine the quarterly frequency equivalent as a robustness test.

8. Flow variables are typically interpolated such that the sum or mean of the monthly interpolated values correspond to the quarterly observed value (Chipman & Lapham, 1995, p. 89) and the sum is typically chosen for flow variables, such as GDP or national income (ibid., p. 92). However, whether the mean or the sum is chosen for the interpolation of GDP levels is not directly relevant because the two can be shown to yield the same logarithmic growth rate, $Y_t^{(h)} = \ln(y_t) - \ln(y_{t-h})$ with y_t denoting the level of GDP in period t (cf. e.g. Siliverstovs & Wochner, 2021; Stock & Watson, 2012; McCracken & Ng, 2019b).

9. Notice that we assume the recession probability index to be released together with FRED-MD variables. The robustness of this assumption will be assessed in Section 4.3.8.

10. For computational reasons, DFFs are based on $B = 200$ bootstrapped samples.

11. As previously explained, we use h-period differenced growth rates, $Y_t^{(h)} = \ln(y_t) - \ln(y_{t-h})$ as dependent

variables, where y_t is the level of GDP in period t (cf. Stock & Watson, 2012; Siliverstovs & Wochner, 2021; also see McCracken & Ng, 2019b). While the quarterly growth rates, $Y_t^{(3)}$, can be simply derived from the interpolated GDP levels via DCO and CL3, this task is more complex for CLU as it does not impute the levels, y_t, but monthly growth rates, $Y_t^{(1)}$. Hence, for $h = 3$, we will reconstruct the quarterly growth rates, $Y_t^{(3)}$, from the monthly growth rates, $Y_t^{(1)}$, by summing over the three adjacent values of $Y_t^{(1)}$, because

$$\begin{aligned} Y_t^{(3)} &= Y_t^{(1)} + Y_{t-1}^{(1)} + Y_{t-2}^{(1)} \\ &= \big(\ln(y_t) - \ln(y_{t-1})\big) + \big(\ln(y_{t-1}) - \ln(y_{t-2})\big) + \big(\ln(y_{t-2}) - \ln(y_{t-3})\big) \\ &= \ln(y_t) - \ln(y_{t-3}). \end{aligned}$$

12. The Chow-Lin (1971) interpolation "CL3" is obtained via the three co-integrated series with FRED-mnemonics `DPCERA3M086SBEA`, `SRVPRD` and `CE16OV`. Stationarity of the errors is warranted in that the (augmented) Dickey-Fuller test (Trapletti, Hornik, & LeBaron, 2019) for the residuals of the regression in levels rejects the null of non-stationarity at the 1% significance level (see e.g. Hill et al., 2011, Chapter 12, for an elaborate discussion of cointegration).

13. For example, for vintage dates in Jan, Feb and May 2018 (Q1 2018), the index is available until Sep 2017 (Q3 2017). Likewise, for vintage dates in April, May and June 2018 (Q2 2018), the index is available until Dec 2017 (Q4 2017) (see Hamilton, 2018).

14. Formally, the regime-switching model is given as (see K. Kim & Swanson, 2016, for a related specification),

$$\hat{Z}_{t+h}^{(1)} = \mathbb{1}(\hat{z}_{t+h} < \varsigma)\hat{Z}_{\text{DL},t+h}^{(1)} + \big(1 - \mathbb{1}(\hat{z}_{t+h} < \varsigma)\big)\hat{Z}_{\text{BETA},t+h}^{(1)},$$

where $\hat{Z}_{\text{DL},t+h}^{(1)}$ designates the prediction from a distributed lag model with BIC chosen number of $Y_{t-i_Y}^{(1)}$ lags with $i_Y \in \{0, 1, \ldots, 4\}$ (cf. Hill et al., 2011) and $\hat{Z}_{\text{BETA},t+h}^{(1)}$ is the predicted value of a beta-regression with a logit-link function for the mean and precision equation (see Cribari-Neto & Zeileis, 2010, and references therein). Both mean and precision equations are BIC optimized over $Y_{t-i_Y}^{(1)}$ lags with $i_Y \in \{0, 1, \ldots, 4\}$, the first i_F linear factors $F_{i_F,t}^{(1)}$ with $i_F \in \{1, \ldots, 10\}$ and the first i_Q quadratic factors $\big(F_{i_Q,t}^{(1)}\big)^2$ with $i_Q \in \{0, \ldots, 10\}$ (cf. Bai & Ng, 2008). Finally, $\varsigma = 0.21$ and $\hat{z}_{t+h}$ corresponds to the predicted raw value of the index from the beta regression.

Acknowledgements

We gratefully acknowledge valuable inputs and comments from Jan-Egbert Sturm, Boriss Siliverstovs, Massimiliano Marcellino, Michael W. McCracken, Philippe Goulet Coulombe, Samad Sarferaz, Alexander Rathke, Philipp Baumann, Petra Ehmann and Oliver Müller. The paper also benefitted from discussions with participants at the SkiLLS Seminar in Engelberg (Switzerland), the research seminar at the Latvian Bank in Riga (Latvia) and the Vienna Workshop on Economic Forecasting (Austria). A previous version of this paper was published under the same title in the KOF Working Paper Series of ETH Zurich in 2020 as well as under the title "Dynamic Factor Forests – A Theory-led Machine Learning Framework for Non-Linear and State-Dependent Short-Term U.S. GDP Growth Predictions" in Ehmann (2021). As outlined in the text, parts of this work (incl. R code) follow and build upon previous work in Ehmann (2020), Wochner (2018) and Siliverstovs and Wochner (2021).

References

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., & Veronese, G. (2010). New Eurocoin: Tracking Economic Growth in Real Time. *The Review of Economics and Statistics*, *92*(4), 1024–1034.

Andrews, D. W. (1993). Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica*, *61*(4), 821–856.

Angelini, E., Henry, J., & Marcellino, M. (2006). Interpolation and Backdating with a Large Information Set. *Journal of Economic Dynamics and Control*, *30*(12), 2693–2724.

Appenzeller, T. (2017). The Scientists' Apprentice. *Science*, *357*(6346), 16–17.

Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-time Measurement of Business Conditions. *Journal of Business & Economic Statistics*, *27*(4), 417–427.

Athey, S., & Imbens, G. W. (2006). Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica*, *74*(2), 431–497.

Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods that Economists Should Know About. *Annual Review of Economics*, *11*, 685–725.

Audrino, F. (2006). Tree-Structured Multiple Regimes in Interest Rates. *Journal of Business & Economic Statistics*, *24*(3), 338–353.

Audrino, F., & Medeiros, M. C. (2011). Modeling and Forecasting Short-Term Interest Rates: The Benefits of Smooth Regimes, Macroeconomic Variables, and Bagging. *Journal of Applied Econometrics*, *26*(6), 999–1022.

Bai, J., & Han, X. (2016). Structural Changes in High Dimensional Factor Models. *Frontiers of Economics in China*, *11*(1), 9–39.

Bai, J., & Ng, S. (2007). Determining the Number of Primitive Shocks in Factor Models. *Journal of Business & Economic Statistics*, *25*(1), 52–60.

Bai, J., & Ng, S. (2008). Forecasting Economic Time Series Using Targeted Predictors. *Journal of Econometrics*, *146*(2), 304–317.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, *101*(473), 119–137.

Balcilar, M. (2019). *mFilter: Miscellaneous Time Series Filters.* CRAN: R Package, Reference Manual. Last accessed on 2022-04-02, from https://cran.r-project.org/package=mFilter.

Bates, B. J., Plagborg-Møller, M., Stock, J. H., & Watson, M. W. (2013). Consistent Factor Estimation in Dynamic Factor Models with Structural Instability. *Journal of Econometrics*, *177*(2), 289–304.

Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. *Test*, *25*(2), 197–227.

Boivin, J., & Ng, S. (2006). Are More Data Always Better for Factor Analysis? *Journal of Econometrics*, *132*(1), 169–194.

Breiman, L. (1996a). Bagging Predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (1996b). Heuristics of Instability and Stabilization in Model Selection. *Annals of Statistics*, *24*(6), 2350–2383.

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees.* Boca Raton, FL: Chapman & Hall/CRC.

Breitung, J., & Eickmeier, S. (2011). Testing for Structural Breaks in Dynamic Factor Models. *Journal of Econometrics*, *163*(1), 71–84.

Bühlmann, P., & Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics*, *30*(4), 927–961.

Bulligan, G., Marcellino, M., & Venditti, F. (2015). Forecasting Economic Activity with Targeted Predictors. *International Journal of Forecasting*, *31*(1), 188–206.

Camacho, M., Perez-Quiros, G., & Poncela, P. (2018). Markov-Switching Dynamic Factor Models in Real Time. *International Journal of Forecasting*, *34*(4), 598–611.

Canty, A. J. (2002). Resampling Methods in R: The boot Package. *R News*, *2*(3), 2–7.

Carstensen, K., Heinrich, M., Reif, M., & Wolters, M. H. (2020). Predicting Ordinary and Severe Recessions with a Three-State Markov-Switching Dynamic Factor Model: An Application to the German Business Cycle. *International Journal of Forecasting*, *36*(3), 829–850.

Chauvet, M., & Hamilton, J. D. (2006). Dating Business Cycle Turning Points. *Contributions to Economic Analysis*, *276*, 1–54.

Chauvet, M., & Potter, S. (2013). Forecasting Output. In G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, p. 141-194). Amsterdam, Netherlands: Elsevier.

Chen, J. C., Dunn, A., Hood, K. K., Driessen, A., & Batch, A. (2019). Off to the Races: A Comparison of Machine Learning and Alternative Data for Predicting Economic Indicators. In K. G. Abraham, R. S. Jarmin, B. Moyer, & M. D. Shapiro (Eds.), *Big Data for 21st Century Economic Statistics.* Chicago, IL: University of Chicago Press.

Chen, X., & Ishwaran, H. (2012). Random Forests for Genomic Data Analysis. *Genomics*, *99*(6), 323–329.

Chipman, J. S., & Lapham, B. J. (1995). Interpolation of Economic Time Series, with Application to German and Swedish Data. In T. Url & A. Wörgötter (Eds.), *Econometrics of Short and Unreliable Time Series* (pp. 89–139). Heidelberg, Germany: Springer.

Chow, G. C., & Lin, A.-l. (1971). Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. *The Review of Economics and Statistics*, *53*(4), 372–375.

Corradi, V., & Swanson, N. R. (2014). Testing for Structural Instability of Factor Augmented Forecasting Models. *Journal of Econometrics*, *182*(1), 100-118.

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, *34*(2), 1–24.

D'Agostino, A., Giannone, D., Lenza, M., & Modugno, M. (2016). Nowcasting Business Cycles: A Bayesian Approach to Dynamic Heterogeneous Factor Models. In E. Hillebrand & S. J. Koopman (Eds.), *Dynamic Factor Models* (Vol. 35, pp. 569–594). Bingley, UK: Emerald Group Publishing Limited.

Dagum, E. B., & Cholette, P. A. (2006). *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series.* New York, NY: Springer Science & Business Media.

Da Rosa, J. C., Veiga, A., & Medeiros, M. C. (2008). Tree-Structured Smooth Transition Regression Models. *Computational Statistics & Data Analysis*,

52(5), 2469–2488.

Denton, F. T. (1971). Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization. *Journal of the American Statistical Association*, *66*(333), 99–102.

Diebold, F. X. (2003). Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting. In M. Dewatripont, L. Hansen, & S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society* (Vol. 3, pp. 115–122). Cambridge, UK: Cambridge University Press.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Diebold, F. X., & Rudebusch, G. D. (1996). Measuring Business Cycles: A Modern Perspective. *The Review of Economics and Statistics*, *78*(1), 67–77.

Döpke, J., Fritsche, U., & Pierdzioch, C. (2017). Predicting Recessions with Boosted Regression Trees. *International Journal of Forecasting*, *33*(4), 745–759.

Doz, C., Ferrara, L., & Pionnier, P.-A. (2020). *Business Cycle Dynamics after the Great Recession: An Extended Markov-Switching Dynamic Factor Model.* Paris, France: OECD Publishing, OECD Statistics Working Papers, No. 2020/01.

Doz, C., & Fuleky, P. (2020). Dynamic Factor Models. In P. Fuleky (Ed.), *Macroeconomic Forecasting in the Era of Big Data* (pp. 27–64). Berlin, Germany: Springer.

Ehmann, D. S. (2020). *Triple Targeted Forests: A Forecasting Experiment for Short-Term Predictions of U.S. GDP Growth in Mixed-Frequency and High-Dimensional Environments.* Unpublished Manuscript.

Ehmann, D. S. (2021). *Essays on Economic Forecasting with Machine Learning.* Zurich, Switzerland: ETH Zurich.

Elliott, G., & Timmermann, A. (2016). *Economic Forecasting.* Princeton, NY: Princeton University Press.

Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., & Aravkin, A. Y. (2020). Sparse Principal Component Analysis via Variable Projection. *SIAM Journal on Applied Mathematics*, *80*(2), 977–1002.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, *542*(7639), 115.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern recognition letters*, *27*(8), 861–874.

Forni, M., Giannone, D., Lippi, M., & Reichlin, L. (2009). Opening the Black Box: Structural Factor Models with Large Cross Sections. *Econometric Theory*, *25*(5), 1319–1347.

Foroni, C., & Marcellino, M. (2013). *A Survey of Econometric Methods for Mixed-Frequency Data.* European University Institute, EUI Working Papers, ECO 2013/02.

Fossati, S. (2018). *A Test for State-Dependent Predictive Ability based on a Markov-Switching Framework (Working Paper).* Edmonton, Alberta, Canada: University of Alberta. Last accessed on 2021-02-12, from https://sites.ualberta.ca/~sfossati/files/pdfs/sd_predict.pdf.

FRBP. (2018). *Survey of Professional Forecasters: Documentation* (Tech. Rep.). Philadelphia, PA: Federal Reserve Bank of Philadelphia. Retrieved 04-11-2019, from https://www.philadelphiafed.org/-/media/research-and-data/real-time-center/survey-of-professional-forecasters/spf-documentation.pdf?la=en

FRBP. (2019). *Probability Variables: Survey of Professional Forecasters.* Philadelphia, PA: Federal Reserve Bank of Philadelphia. Retrieved 04-11-2019, from https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/historical-data/probability-variables

FRED. (2019). GDP-Based Recession Indicator Index. St. Louis, MO: Federal Reserve Bank of St. Louis, Federal Reserve Economic Data (FRED). Retrieved 24-05-2018, from https://fred.stlouisfed.org/series/JHGDPBRINDX (FRED Mnemonic: JHGDPBRINDX)

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189–1232.

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Garcia, M. G., Medeiros, M. C., & Vasconcelos, G. F. (2017). Real-Time Inflation Forecasting with High-Dimensional Models: The Case of Brazil. *International Journal of Forecasting*, *33*(3), 679–693.

Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random Forest Methodology for Model-based Recursive Partitioning: the mobForest Package for R. *BMC Bioinformatics*, *14*(1), 125.

Giannone, D., Reichlin, L., & Sala, L. (2006). VARs, Common Factors and the Empirical Validation of Equilibrium Business Cycle Models. *Journal of Econometrics*, *132*(1), 257–279.

Goulet Coulombe, P. (2020). *The Macroeconomy as a Random Forest.* Last accessed on 2021-02-11, from https://arxiv.org/abs/2006.12724.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2019). *How is Machine Learning Useful for Macroeconomic Forecasting?* CIRANO Working Papers 2019s-22.

Guérin, P., & Marcellino, M. (2013). Markov-Switching MIDAS Models. *Journal of Business & Economic Statistics*, *31*(1), 45–56.

Hamilton, J. D. (2018). *The Econbrowser Recession Indicator Index.* Retrieved 10-12-2018, from http://econbrowser.com/recession-index

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York, NY: Springer.

Hill, R. C., Griffiths, W. E., Lim, G. C., & Lim, M. A. (2011). *Principles of Econometrics* (4th ed.). Hoboken, NJ: Wiley & Sons, Inc.

Hodrick, R. J., & Prescott, E. C. (1997). Postwar US Business Cycles: An

Empirical Investigation. *Journal of Money, credit, and Banking*, 1–16.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67.

Hothorn, T., & Zeileis, A. (2020). *partykit: A Toolkit for Recursive Partytioning.* CRAN: R Package, Reference Manual. Last accessed on 2020-11-24, from https://cran.r-project.org/package=partykit.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., ... Razbash, S. (2019). *forecast: Forecasting Functions for Time Series and Linear Models.* CRAN: R Package, Reference Manual. Last accessed on 2021-02-21, from https://cran.r-project.org/package=forecast.

Hyndman, R. J., & Koehler, A. B. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York, NY: Springer.

Jones, P. J., Mair, P., Simon, T., & Zeileis, A. (2020). Network Trees: A Method for Recursively Partitioning Covariance Structures. *Psychometrika*, *85*, 926–945.

Khaidem, L., Saha, S., & Dey, S. R. (2016). *Predicting the Direction of Stock Market Prices Using Random Forest.* Last accessed on 2021-02-12, from https://arxiv.org/abs/1605.00003.

Kim, H. H., & Swanson, N. R. (2014). Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence. *Journal of Econometrics*, *178*, 352–367.

Kim, H. H., & Swanson, N. R. (2018). Mining Big Data using Parsimonious Factor, Machine Learning, Variable Selection and Shrinkage Methods. *International Journal of Forecasting*, *34*(2), 339–354.

Kim, K., & Swanson, N. R. (2016). *Mixing Mixed Frequency and Diffusion Indices in Good Times and in Bad.* Last accessed on 2021-02-12, from https://ssrn.com/abstract=2998272.

Kock, A. B., & Teräsvirta, T. (2011). Forecasting with Nonlinear Time Series Models. In M. P. Clements & D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting* (pp. 61–87). Oxford, UK: Oxford University Press.

Kopf, J., Augustin, T., & Strobl, C. (2013). The Potential of Model-Based Recursive Partitioning in the Social Sciences: Revisiting Ockham's Razor. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences* (pp. 97–117). Abingdon Oxfordshire, UK: Routledge Taylor & Francis.

Korobilis, D. (2017). *Forecasting with Many Predictors using Message Passing Algorithms.* University of Essex, Essex Finance Centre Working Paper Series, Working Paper No18: 04-2017.

Kuan, C.-M. (2002). *Lecture on the Markov Switching Model* (Tech. Rep.). Taipei, Taiwan: Academia Sinica, Institute of Economics.

Kuhn, M. (2008). Building Predictive Models in R using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26.

Kuhn, M. (2019). *The caret package.* GitHub: Caret Package. Last accessed on 2022-04-03, from https://topepo.github.io/caret/.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436.

Leiva-Leon, D., Pérez-Quirós, G., & Rots, E. (2020). *Real-time Weakness of the Global Economy: A First Assessment of the Coronavirus Crisis.* European Central Bank, ECB Working Paper, No. 2381.

McAfee, A., & Brynjolfsson, E. (2017). *Machine, Platform, Crowd: Harnessing our Digital Future.* New York, NY: WW Norton & Company.

McCracken, M. W. (2007). Asymptotics for Out of Sample Tests of Granger Causality. *Journal of Econometrics*, *140*(2), 719–752.

McCracken, M. W. (2019). *FRED-MD: A Monthly Database for Macroeconomic Research.* St. Louis, MO: Federal Reserve Bank of St. Louis. Last accessed on 2021-02-12, from https://research.stlouisfed.org/econ/mccracken/fred-databases/.

McCracken, M. W., & Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

McCracken, M. W., & Ng, S. (2019a). *FRED-MD: A Monthly Database for Macroeconomic Research.* St. Louis, MO: Federal Reserve Bank of St. Louis. Retrieved 12-03-2019, from https://s3.amazonaws.com/files.fred.stlouisfed.org/fred-md/Appendix_Tables_Update.pdf (Online Appendix for FRED-MD)

McCracken, M. W., & Ng, S. (2019b). *FRED-QD: A Quarterly Database for Macroeconomic Research.* St. Louis, MO: Federal Reserve Bank of St. Louis. Retrieved 12-03-2019, from https://s3.amazonaws.com/files.fred.stlouisfed.org/fred-md/FRED-QD_appendix.pdf (Online Appendix for FRED-QD)

McCracken, M. W., & Ng, S. (2020). *FRED-QD: A Quarterly Database for Macroeconomic Research.* National Bureau of Economic Research, NBER Working Paper, No. 26872.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, *39*(1), 98-119.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, *3*(2), 1–21.

Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, *22*(2), 303–341.

Mukherjee, S. (2017). AI versus MD – What Happens when Diagnosis is Automated? *The New Yorker*. April 3, 2017 issue. Last accessed on 2021-02-12, from https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

NBER. (2020). *US Business Cycle Expansions and Contractions.* Cambridge, MA: National Bureau of Economic Research (NBER). Last accessed on 2020-04-03, from http://www.nber.org/cycles.

NBER. (2021). *US Business Cycle Expansions and Contractions.* Cambridge, MA: National Bureau of Economic Research (NBER). Last accessed on 2021-01-03, from https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions.

Ng, S. (2014). Boosting Recessions. *Canadian Journal of Economics/Revue Canadienne d'Économique*, *47*(1), 1–34.

Pierdzioch, C., & Gupta, R. (2020). Uncertainty and Forecasts of US Recessions. *Studies in Nonlinear Dynamics & Econometrics*, *24*(4).

Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313.

Quan, Z., Wang, Z., Gan, G., & Valdez, E. A. (2020). *Hybrid Tree-based Models for Insurance Claims.* Last accessed on 2021-02-12, from https://arxiv.org/abs/2006.05617.

Racine, J. (2000). Consistent Cross-Validatory Model-Selection for Dependent Data: hv-Block Cross-Validation. *Journal of Econometrics*, *99*(1), 39–61.

Rossi, B. (2013). Advances in Forecasting under Instability. In G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, pp. 1203–1324). Amsterdam, Netherlands: Elsevier.

Sax, C., & Steiner, P. (2013). Temporal Disaggregation of Time Series. *The R Journal*, *5*(2), 80–87.

Sax, Christopher and Steiner, Peter and Di Fonzo, Tommaso. (2020). *tempdisagg: Methods for Temporal Disaggregation and Interpolation of Time Series.* CRAN: R Package, Reference Manual. Last accessed on 2020-12-15, from https://cran.r-project.org/package=tempdisagg.

Schlosser, L., Hothorn, T., Stauffer, R., & Zeileis, A. (2019). Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain. *The Annals of Applied Statistics*, *13*(3), 1564–1589.

Schölkopf, B., Smola, A., & Müller, K.-R. (1997). Kernel Principal Component Analysis. In W. Gerstner, A. Germond, M. Hasler, & J.-D. Nicoud (Eds.), *Artificial Neural Networks — ICANN'97* (pp. 583–588). Germany, Berlin: Springer.

Seibold, H., Zeileis, A., & Hothorn, T. (2016). Model-based Recursive Partitioning for Subgroup Analyses. *The International Journal of Biostatistics*, *12*(1), 45–63.

Siliverstovs, B. (2017a). Dissecting Models' Forecasting Performance. *Economic Modelling*, *67*, 294–299.

Siliverstovs, B. (2017b). International Stock Return Predictability: On the Role of the United States in Bad and Good Times. *Applied Economics Letters*, *24*(11), 771–773.

Siliverstovs, B. (2020). Assessing Nowcast Accuracy of US GDP Growth in Real Time: The Role of Booms and Busts. *Empirical Economics*, *58*(1), 7–27.

Siliverstovs, B., & Wochner, D. S. (2021). State-Dependent Evaluation of Pre-

dictive Ability. *Journal of Forecasting*, *40*(3), 547–574.

Stock, J. H., & Watson, M. (2009). Forecasting in Dynamic Factor Models Subject to Structural Instability. In J. Castle & N. Shephard (Eds.), *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry* (p. 173-205). Oxford, UK: Oxford University Press.

Stock, J. H., & Watson, M. W. (2002a). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2002b). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162.

Stock, J. H., & Watson, M. W. (2006). Forecasting with Many Predictors. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 515–554). Amsterdam, Netherlands: Elsevier.

Stock, J. H., & Watson, M. W. (2011). Dynamic Factor Models. In M. P. Clements & D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting* (pp. 35–59). Oxford, UK: Oxford University Press.

Stock, J. H., & Watson, M. W. (2012). Generalized Shrinkage Methods for Forecasting using Many Predictors. *Journal of Business & Economic Statistics*, *30*(4), 481–493.

Stock, J. H., & Watson, M. W. (2016). Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of Macroeconomics* (Vol. 2, pp. 415–525). Amsterdam, Netherlands: Elsevier.

Stock, J. H., & Watson, M. W. (2017). Twenty Years of Time Series Econometrics in Ten Pictures. *Journal of Economic Perspectives*, *31*(2), 59–86.

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135–153.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Trapletti, A., Hornik, K., & LeBaron, B. (2019). *tseries: Time Series Analysis and Computational Finance.* CRAN: R Package, Reference Manual. Retrieved on 2019-10-28, from https://cran.r-project.org/package=tseries.

Tu, Y., & Lee, T.-H. (2019). Forecasting using Supervised Factor Models. *Journal of Management Science and Engineering*, *4*(1), 12–27.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives*, *28*(2), 3–27.

Veltri, G. A. (2017). Big Data is Not Only About Data: The Two Cultures of Modelling. *Big Data & Society*, *4*(1), 1–6.

Welch, I., & Goyal, A. (2008). A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

Wochner, D. S. (2018). *Reducing Dimensionality for Economic Time Series*

Forecasting: A Double Predictor Targeting Strategy (Unpublished M.Sc. Thesis). Oxford, UK: University of Oxford.

Zeileis, A. (2005). A Unified Approach to Structural Change Tests based on ML Scores, F Statistics, and OLS Residuals. *Econometric Reviews*, *24*(4), 445–466.

Zeileis, A., & Hornik, K. (2007). Generalized M-Fluctuation Tests for Parameter Instability. *Statistica Neerlandica*, *61*(4), 488–508.

Zeileis, A., & Hothorn, T. (2015). Parties, Models, Mobsters: A New Implementation of Model-Based Recursive Partitioning in R. Retrieved 2019-10-15, from https://cran.rstudio.org/web/packages/partykit/vignettes/mob.pdf (R Package vignette)

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.

A. Appendix: Tables and Figures

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
			DFT-NUM(F)	**0.772*****	DFF-NUM(F)	**0.753*****
	AR4		DFT-NUM-PLUS-T(F)	**0.762*****	DFF-NUM-PLUS-T(F)	**0.730*****
	0.924		DFT-NUM-PLUS-X(F)	0.924	DFF-NUM-PLUS-X(F)	**0.799*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
			DFT-NUM(TF)	**0.755*****	DFF-NUM(TF)	**0.730*****
	CADL		DFT-NUM-PLUS-T(TF)	**0.801**	DFF-NUM-PLUS-T(TF)	**0.713*****
	0.922		DFT-NUM-PLUS-X(TF)	0.869	DFF-NUM-PLUS-X(TF)	**0.767*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
			DFT-NUM(F)	**0.844*****	DFF-NUM(F)	**0.816*****
	AR4		DFT-NUM-PLUS-T(F)	**0.852****	DFF-NUM-PLUS-T(F)	**0.777*****
	0.970		DFT-NUM-PLUS-X(F)	**0.897***	DFF-NUM-PLUS-X(F)	**0.868*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
			DFT-NUM(TF)	**0.853****	DFF-NUM(TF)	**0.803*****
	CADL		DFT-NUM-PLUS-T(TF)	**0.806*****	DFF-NUM-PLUS-T(TF)	**0.754*****
	0.943		DFT-NUM-PLUS-X(TF)	0.943	DFF-NUM-PLUS-X(TF)	**0.831*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; no pruning). Lasso-based regularization for the generalized partitioning sets 'PLUS-T' and 'PLUS-X'. The acronyms PLUS-T and PLUS-X refer to sets of splitting variables that incorporate, in addition to the ERI-Index, time (T) and all contemporaneous explanatory variables (X), respectively. For more details, see notes in Table 1.

Table A.1: Extended Results: Lasso-based Penalization with Generalized Partitioning Sets (relative RMSFE)

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
			DFT-NUM-PLS(F)	0.879	DFF-NUM-PLS(F)	**0.828***
	AR4		DFT-NUM-KPCA(F)	**0.794*****	DFF-NUM-KPCA(F)	**0.755*****
	0.924		DFT-NUM-SPCA(F)	**0.816****	DFF-NUM-SPCA(F)	**0.780*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
			DFT-NUM-PLS(TF)	0.819	DFF-NUM-PLS(TF)	**0.739*****
	CADL		DFT-NUM-KPCA(TF)	**0.768*****	DFF-NUM-KPCA(TF)	**0.716*****
	0.922		DFT-NUM-SPCA(TF)	**0.777****	DFF-NUM-SPCA(TF)	**0.732*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
			DFT-NUM-PLS(F)	**0.877****	DFF-NUM-PLS(F)	**0.867****
	AR4		DFT-NUM-KPCA(F)	**0.829*****	DFF-NUM-KPCA(F)	**0.806*****
	0.970		DFT-NUM-SPCA(F)	0.939	DFF-NUM-SPCA(F)	**0.867****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
			DFT-NUM-PLS(TF)	**0.803*****	DFF-NUM-PLS(TF)	**0.781*****
	CADL		DFT-NUM-KPCA(TF)	**0.831*****	DFF-NUM-KPCA(TF)	**0.774*****
	0.943		DFT-NUM-SPCA(TF)	**0.872****	DFF-NUM-SPCA(TF)	**0.816*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The acronym PLS refers to Partial Least Squares and the acronyms KPCA and SPCA refer to Kernel and Sparse PCA, respectively. For more details, see notes in Table 1.

Table A.2: Extended Results: Novel Factors (PLS, KPCA and SPCA) (relative RMSFE)

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
			DFT-PLUS-T-PLS(F)	**0.762*****	DFF-PLUS-T-PLS(F)	**0.774*****
	AR4		DFT-PLUS-T-KPCA(F)	**0.798*****	DFF-PLUS-T-KPCA(F)	**0.751*****
	0.924		DFT-PLUS-T-SPCA(F)	**0.818****	DFF-PLUS-T-SPCA(F)	**0.753*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
			DFT-PLUS-T-PLS(TF)	**0.738*****	DFF-PLUS-T-PLS(TF)	**0.709*****
	CADL		DFT-PLUS-T-KPCA(TF)	**0.775****	DFF-PLUS-T-KPCA(TF)	**0.711*****
	0.922		DFT-PLUS-T-SPCA(TF)	**0.793***	DFF-PLUS-T-SPCA(TF)	**0.721*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
			DFT-PLUS-T-PLS(F)	**0.810*****	DFF-PLUS-T-PLS(F)	**0.837*****
	AR4		DFT-PLUS-T-KPCA(F)	**0.895***	DFF-PLUS-T-KPCA(F)	**0.794*****
	0.970		DFT-PLUS-T-SPCA(F)	**0.884***	DFF-PLUS-T-SPCA(F)	**0.825*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
			DFT-PLUS-T-PLS(TF)	**0.793*****	DFF-PLUS-T-PLS(TF)	**0.763*****
	CADL		DFT-PLUS-T-KPCA(TF)	**0.827****	DFF-PLUS-T-KPCA(TF)	**0.748*****
	0.943		DFT-PLUS-T-SPCA(TF)	**0.891***	DFF-PLUS-T-SPCA(TF)	**0.774*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; no pruning). Lasso-based regularization is applied. For space considerations 'DFT-NUM-PLUS-T' and 'DFF-NUM-PLUS-T' were abbreviated to 'DFT-PLUS-T' and 'DFF-PLUS-T'. The acronym PLS refers to Partial Least Squares and the acronyms KPCA and SPCA refer to Kernel and Sparse PCA, respectively. For more details, see notes in Table 1.

Table A.3: Extended Results: Lasso-based Penalization for Generalized Partitioning Sets with Novel Factors (relative RMSFE)

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1		Factors	OS-DFM(F)	*0.892*	BS-DFM(F)	*0.890*
	HMN		OS-RT(F)	1.182	BS-RF(F)	0.877
	1.063		DFT-NUM(F)	**0.887***	DFF-NUM(F)	**0.844*****
			DFT-BIN50(F)	**0.890**	DFF-BIN50(F)	**0.858*****
			DFT-NUM-PLUS-T(F)	0.902	DFF-NUM-PLUS-T(F)	**0.843*****
	AR4		DFT-NUM-PLUS-X(F)	0.895	DFF-NUM-PLUS-X(F)	**0.872*****
	0.983		DFT-NUM-RIDGE(F)	**0.836*****	DFF-NUM-RIDGE(F)	**0.839*****
			DFT-NUM-LASSO(F)	**0.826*****	DFF-NUM-LASSO(F)	**0.830*****
		Trgt. Fact.	OS-DFM(TF)	*0.867*	BS-DFM(TF)	*0.866*
	ARL		OS-RT(TF)	1.270	BS-RF(TF)	0.867^{+}
	1.003		DFT-NUM(TF)	**0.867**	DFF-NUM(TF)	**0.828*****
			DFT-BIN50(TF)	**0.867**	DFF-BIN50(TF)	**0.843*****
			DFT-NUM-PLUS-T(TF)	**0.865**	DFF-NUM-PLUS-T(TF)	**0.830*****
	CADL		DFT-NUM-PLUS-X(TF)	**0.867**	DFF-NUM-PLUS-X(TF)	**0.854*****
	0.931		DFT-NUM-RIDGE(TF)	**0.834*****	DFF-NUM-RIDGE(TF)	**0.831*****
			DFT-NUM-LASSO(TF)	**0.831*****	DFF-NUM-LASSO(TF)	**0.820*****
h=3		Factors	OS-DFM(F)	*0.883*	BS-DFM(F)	*0.880*
	HMN		OS-RT(F)	1.404	BS-RF(F)	0.937
	1.078		DFT-NUM(F)	**0.779*****	DFF-NUM(F)	**0.762*****
			DFT-BIN50(F)	**0.785*****	DFF-BIN50(F)	**0.783*****
			DFT-NUM-PLUS-T(F)	**0.757*****	DFF-NUM-PLUS-T(F)	**0.781*****
	AR4		DFT-NUM-PLUS-X(F)	**0.796*****	DFF-NUM-PLUS-X(F)	**0.807*****
	0.989		DFT-NUM-RIDGE(F)	**0.789*****	DFF-NUM-RIDGE(F)	**0.778*****
			DFT-NUM-LASSO(F)	**0.779*****	DFF-NUM-LASSO(F)	**0.769*****
		Trgt. Fact.	OS-DFM(TF)	*0.864*	BS-DFM(TF)	*0.862*
	ARL		OS-RT(TF)	1.440	BS-RF(TF)	0.927
	1.010		DFT-NUM(TF)	**0.782*****	DFF-NUM(TF)	**0.751*****
			DFT-BIN50(TF)	**0.788*****	DFF-BIN50(TF)	**0.780*****
			DFT-NUM-PLUS-T(TF)	**0.772*****	DFF-NUM-PLUS-T(TF)	**0.771*****
	CADL		DFT-NUM-PLUS-X(TF)	0.920	DFF-NUM-PLUS-X(TF)	**0.805*****
	0.932		DFT-NUM-RIDGE(TF)	**0.783*****	DFF-NUM-RIDGE(TF)	**0.768*****
			DFT-NUM-LASSO(TF)	**0.774*****	DFF-NUM-LASSO(TF)	**0.753*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; Chow-Lin (CLU) interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The penalized DFTs and DFFs are unpruned. For more details, see notes in Table 1.

Table A.4: Robustness Results: Chow-Lin (CLU) Interpolation

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.943*	BS-DFM(F)	*0.942*
			OS-RT(F)	1.451	BS-RF(F)	0.941
	1.004		DFT-NUM(F)	**0.943**	DFF-NUM(F)	**0.922*****
			DFT-BIN50(F)	**0.943**	DFF-BIN50(F)	**0.935*****
			DFT-NUM-PLUS-T(F)	0.955	DFF-NUM-PLUS-T(F)	**0.929*****
	AR4		DFT-NUM-PLUS-X(F)	0.965	DFF-NUM-PLUS-X(F)	**0.941***
	0.990		DFT-NUM-RIDGE(F)	0.963	DFF-NUM-RIDGE(F)	**0.932****
			DFT-NUM-LASSO(F)	0.972	DFF-NUM-LASSO(F)	**0.932****
		Trgt. Fact.	OS-DFM(TF)	*0.927*	BS-DFM(TF)	*0.926*
	ARL		OS-RT(TF)	1.454	BS-RF(TF)	0.951
	0.998		DFT-NUM(TF)	**0.927**	DFF-NUM(TF)	**0.908*****
			DFT-BIN50(TF)	**0.927**	DFF-BIN50(TF)	**0.922*****
			DFT-NUM-PLUS-T(TF)	0.949	DFF-NUM-PLUS-T(TF)	**0.910*****
	CADL		DFT-NUM-PLUS-X(TF)	0.929	DFF-NUM-PLUS-X(TF)	0.934
	0.971		DFT-NUM-RIDGE(TF)	0.949	DFF-NUM-RIDGE(TF)	**0.925***
			DFT-NUM-LASSO(TF)	0.944	DFF-NUM-LASSO(TF)	**0.910****
h=3	HMN	Factors	OS-DFM(F)	*0.930*	BS-DFM(F)	*0.928*
			OS-RT(F)	1.563	BS-RF(F)	0.969
	1.014		DFT-NUM(F)	**0.907****	DFF-NUM(F)	**0.861*****
			DFT-BIN50(F)	**0.918***	DFF-BIN50(F)	**0.876*****
			DFT-NUM-PLUS-T(F)	**0.897****	DFF-NUM-PLUS-T(F)	**0.841*****
	AR4		DFT-NUM-PLUS-X(F)	0.966	DFF-NUM-PLUS-X(F)	**0.889*****
	0.980		DFT-NUM-RIDGE(F)	**0.897****	DFF-NUM-RIDGE(F)	**0.866*****
			DFT-NUM-LASSO(F)	**0.890****	DFF-NUM-LASSO(F)	**0.864*****
		Trgt. Fact.	OS-DFM(TF)	*0.931*	BS-DFM(TF)	*0.930*
	ARL		OS-RT(TF)	1.474	BS-RF(TF)	0.962
	1.012		DFT-NUM(TF)	**0.897****	DFF-NUM(TF)	**0.864*****
			DFT-BIN50(TF)	0.938	DFF-BIN50(TF)	**0.890*****
			DFT-NUM-PLUS-T(TF)	**0.916***	DFF-NUM-PLUS-T(TF)	**0.826*****
	CADL		DFT-NUM-PLUS-X(TF)	1.007	DFF-NUM-PLUS-X(TF)	**0.888*****
	0.956		DFT-NUM-RIDGE(TF)	**0.889****	DFF-NUM-RIDGE(TF)	**0.866*****
			DFT-NUM-LASSO(TF)	**0.859*****	DFF-NUM-LASSO(TF)	**0.844*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; Chow-Lin (CL3) interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The penalized DFTs and DFFs are unpruned. For more details, see notes in Table 1.

Table A.5: Robustness Results: Chow-Lin (CL3) Interpolation

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
			OS-DFM(F)	*0.885*	BS-DFM(F)	*0.881*
	HMN		OS-RT(F)	1.182	BS-RF(F)	0.877
	0.998	Factors	DFT-NUM(F)	**0.809*****	DFF-NUM(F)	**0.772*****
			DFT-BIN50(F)	**0.811*****	DFF-BIN50(F)	**0.792*****
			DFT-NUM-PLUS-T(F)	**0.806*****	DFF-NUM-PLUS-T(F)	**0.780*****
	AR4		DFT-NUM-PLUS-X(F)	**0.822*****	DFF-NUM-PLUS-X(F)	**0.815*****
	0.935		DFT-NUM-RIDGE(F)	**0.802*****	DFF-NUM-RIDGE(F)	**0.776*****
h=1			DFT-NUM-LASSO(F)	**0.800*****	DFF-NUM-LASSO(F)	**0.770*****
			OS-DFM(TF)	*0.814*	BS-DFM(TF)	*0.813*
	ARL		OS-RT(TF)	1.208	BS-RF(TF)	0.879
	0.909	Trgt. Fact.	DFT-NUM(TF)	**0.742*****	DFF-NUM(TF)	**0.715*****
			DFT-BIN50(TF)	**0.767*****	DFF-BIN50(TF)	**0.740*****
			DFT-NUM-PLUS-T(TF)	**0.763*****	DFF-NUM-PLUS-T(TF)	**0.730*****
	CADL		DFT-NUM-PLUS-X(TF)	**0.765*****	DFF-NUM-PLUS-X(TF)	**0.746*****
	0.938		DFT-NUM-RIDGE(TF)	**0.760*****	DFF-NUM-RIDGE(TF)	**0.727*****
			DFT-NUM-LASSO(TF)	**0.753*****	DFF-NUM-LASSO(TF)	**0.727*****
			OS-DFM(F)	*0.964*	BS-DFM(F)	*0.956*
	HMN		OS-RT(F)	1.209	BS-RF(F)	0.971
	1.020	Factors	DFT-NUM(F)	**0.839*****	DFF-NUM(F)	**0.805*****
			DFT-BIN50(F)	**0.853*****	DFF-BIN50(F)	**0.841*****
			DFT-NUM-PLUS-T(F)	**0.833*****	DFF-NUM-PLUS-T(F)	**0.848*****
	AR4		DFT-NUM-PLUS-X(F)	**0.855*****	DFF-NUM-PLUS-X(F)	**0.899*****
	0.973		DFT-NUM-RIDGE(F)	**0.837*****	DFF-NUM-RIDGE(F)	**0.805*****
h=3			DFT-NUM-LASSO(F)	**0.839*****	DFF-NUM-LASSO(F)	**0.810*****
			OS-DFM(TF)	*0.940*	BS-DFM(TF)	*0.933*
	ARL		OS-RT(TF)	1.422	BS-RF(TF)	0.970
	0.980	Trgt. Fact.	DFT-NUM(TF)	**0.775*****	DFF-NUM(TF)	**0.753*****
			DFT-BIN50(TF)	**0.850*****	DFF-BIN50(TF)	**0.819*****
			DFT-NUM-PLUS-T(TF)	**0.797*****	DFF-NUM-PLUS-T(TF)	**0.819*****
	CADL		DFT-NUM-PLUS-X(TF)	**0.820*****	DFF-NUM-PLUS-X(TF)	**0.847*****
	0.960		DFT-NUM-RIDGE(TF)	**0.786*****	DFF-NUM-RIDGE(TF)	**0.761*****
			DFT-NUM-LASSO(TF)	**0.790*****	DFF-NUM-LASSO(TF)	**0.764*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: rolling scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The rolling window has a length of 300 observations. The penalized DFTs and DFFs are unpruned. For more details, see notes in Table 1.

Table A.6: Robustness Results: Rolling Windows

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1		Factors	OS-DFM(F)	*0.865*	BS-DFM(F)	*0.864*
	HMN		OS-RT(F)	1.325	BS-RF(F)	0.878
	0.998		DFT-NUM(F)	**0.819*****	DFF-NUM(F)	**0.797*****
			DFT-BIN50(F)	**0.834*****	DFF-BIN50(F)	**0.816*****
			DFT-NUM-PLUS-T(F)	**0.859*****	DFF-NUM-PLUS-T(F)	**0.791*****
	AR4		DFT-NUM-PLUS-X(F)	0.867	DFF-NUM-PLUS-X(F)	**0.828*****
	0.927		DFT-NUM-RIDGE(F)	**0.826*****	DFF-NUM-RIDGE(F)	**0.802*****
			DFT-NUM-LASSO(F)	**0.828*****	DFF-NUM-LASSO(F)	**0.798*****
		Trgt. Fact.	OS-DFM(TF)	*0.831*	BS-DFM(TF)	*0.830*
	ARL		OS-RT(TF)	1.417	BS-RF(TF)	0.872
	0.894		DFT-NUM(TF)	**0.811*****	DFF-NUM(TF)	**0.773*****
			DFT-BIN50(TF)	**0.815*****	DFF-BIN50(TF)	**0.793*****
			DFT-NUM-PLUS-T(TF)	0.861	DFF-NUM-PLUS-T(TF)	**0.767*****
	CADL		DFT-NUM-PLUS-X(TF)	**0.819*****	DFF-NUM-PLUS-X(TF)	**0.799*****
	0.924		DFT-NUM-RIDGE(TF)	**0.811*****	DFF-NUM-RIDGE(TF)	**0.783*****
			DFT-NUM-LASSO(TF)	**0.807*****	DFF-NUM-LASSO(TF)	**0.779*****
h=3		Factors	OS-DFM(F)	*0.947*	BS-DFM(F)	*0.943*
	HMN		OS-RT(F)	1.565	BS-RF(F)	0.991
	1.015		DFT-NUM(F)	**0.900*****	DFF-NUM(F)	**0.859*****
			DFT-BIN50(F)	**0.902*****	DFF-BIN50(F)	**0.880*****
			DFT-NUM-PLUS-T(F)	1.027	DFF-NUM-PLUS-T(F)	**0.853*****
	AR4		DFT-NUM-PLUS-X(F)	**0.941****	DFF-NUM-PLUS-X(F)	**0.900*****
	0.975		DFT-NUM-RIDGE(F)	**0.896*****	DFF-NUM-RIDGE(F)	**0.865*****
			DFT-NUM-LASSO(F)	**0.892*****	DFF-NUM-LASSO(F)	**0.863*****
		Trgt. Fact.	OS-DFM(TF)	*0.931*	BS-DFM(TF)	*0.926*
	ARL		OS-RT(TF)	1.515	BS-RF(TF)	0.987
	0.987		DFT-NUM(TF)	**0.906*****	DFF-NUM(TF)	**0.841*****
			DFT-BIN50(TF)	**0.908*****	DFF-BIN50(TF)	**0.871*****
			DFT-NUM-PLUS-T(TF)	0.935**	DFF-NUM-PLUS-T(TF)	**0.824*****
	CADL		DFT-NUM-PLUS-X(TF)	0.972	DFF-NUM-PLUS-X(TF)	**0.878*****
	0.944		DFT-NUM-RIDGE(TF)	**0.898*****	DFF-NUM-RIDGE(TF)	**0.859*****
			DFT-NUM-LASSO(TF)	**0.887*****	DFF-NUM-LASSO(TF)	**0.852*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1985 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The penalized DFTs and DFFs are unpruned. For more details, see notes in Table 1.

Table A.7: Robustness Results: Extended Forecasting Window (1985-2018)

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=0	HMN	Factors	OS-DFM(F)	*0.739*	BS-DFM(F)	*0.744*
	1.115		OS-RT(F)	1.335	BS-RF(F)	0.859
	AR4		DFT-NUM(F)	**0.709****	DFF-NUM(F)	**0.686*****
	1.004		DFT-BIN50(F)	**0.703*****	DFF-BIN50(F)	**0.708*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.754*	BS-DFM(TF)	*0.762*
	1.025		OS-RT(TF)	1.419	BS-RF(TF)	0.847
	CADL		DFT-NUM(TF)	**0.719*****	DFF-NUM(TF)	**0.703*****
	0.869		DFT-BIN50(TF)	**0.705*****	DFF-BIN50(TF)	**0.722*****
h=1	HMN	Factors	OS-DFM(F)	*0.973*	BS-DFM(F)	*0.977*
	1.060		OS-RT(F)	1.637	BS-RF(F)	0.967
	AR4		DFT-NUM(F)	**0.899*****	DFF-NUM(F)	**0.903*****
	1.006		DFT-BIN50(F)	**0.920*****	DFF-BIN50(F)	**0.921*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.866*	BS-DFM(TF)	*0.874*
	1.002		OS-RT(TF)	1.501	BS-RF(TF)	0.940
	CADL		DFT-NUM(TF)	**0.825*****	DFF-NUM(TF)	**0.840*****
	0.945		DFT-BIN50(TF)	**0.862***	DFF-BIN50(TF)	**0.863****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; quarterly frequency; Q1 Jan. 1998 first vintage; ERI-IDX partitioning variable; without pruning). The results for the predictions in the 3rd month of every quarter are shown. The DM-tests require $h = 1$ for both forecasting horizons and were derived accordingly (cf. Hyndman et al., 2019). The hyper-parameter for minsize was determined via cross-validation. For more details, see notes in Table 1.

Table A.8: Robustness Results: Quarterly Frequency

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
	AR4		DFT-NUM(F)	**0.770*****	DFF-NUM(F)	**0.748*****
	0.924		DFT-BIN50(F)	**0.763*****	DFF-BIN50(F)	**0.759*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
	CADL		DFT-NUM(TF)	**0.760*****	DFF-NUM(TF)	**0.727*****
	0.922		DFT-BIN50(TF)	**0.739*****	DFF-BIN50(TF)	**0.734*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
	AR4		DFT-NUM(F)	**0.843*****	DFF-NUM(F)	**0.815*****
	0.970		DFT-BIN50(F)	**0.827*****	DFF-BIN50(F)	**0.821*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
	CADL		DFT-NUM(TF)	**0.853****	DFF-NUM(TF)	**0.802*****
	0.943		DFT-BIN50(TF)	**0.819*****	DFF-BIN50(TF)	**0.812*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable). The estimation employs pre-pruning at significance level 0.1%. For more details, see notes in Table 1.

Table A.9: Robustness Results: Pre-Pruning

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.873*	BS-DFM(F)	*0.872*
	1.001		OS-RT(F)	1.388	BS-RF(F)	0.855
	AR4		DFT-NUM(F)	**0.801*****	DFF-NUM(F)	**0.777*****
	0.924		DFT-BIN50(F)	**0.791*****	DFF-BIN50(F)	**0.788*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.895*	BS-DFM(TF)	*0.896*
	0.894		OS-RT(TF)	1.222	BS-RF(TF)	0.854
	CADL		DFT-NUM(TF)	**0.804*****	DFF-NUM(TF)	**0.795*****
	0.922		DFT-BIN50(TF)	**0.808*****	DFF-BIN50(TF)	**0.806*****
h=3	HMN	Factors	OS-DFM(F)	*0.897*	BS-DFM(F)	*0.879*
	1.014		OS-RT(F)	1.434	BS-RF(F)	0.953
	AR4		DFT-NUM(F)	**0.802*****	DFF-NUM(F)	**0.747*****
	0.970		DFT-BIN50(F)	**0.796****	DFF-BIN50(F)	**0.793****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.915*	BS-DFM(TF)	*0.908*
	0.975		OS-RT(TF)	1.405	BS-RF(TF)	0.951
	CADL		DFT-NUM(TF)	**0.785*****	DFF-NUM(TF)	**0.759*****
	0.943		DFT-BIN50(TF)	**0.811****	DFF-BIN50(TF)	**0.810****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The RTs, RFs, DFMs, DFTs and DFFs were estimated with 1 factor. For more details, see notes in Table 1.

Table A.10: Robustness Results: One Factor

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.767*	BS-DFM(F)	*0.765*
	1.001		OS-RT(F)	1.317	BS-RF(F)	0.853
	AR4		DFT-NUM(F)	**0.767**	DFF-NUM(F)	**0.714*****
	0.924		DFT-BIN50(F)	**0.767**	DFF-BIN50(F)	**0.728*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.772*	BS-DFM(TF)	*0.773*
	0.894		OS-RT(TF)	1.196	BS-RF(TF)	0.850
	CADL		DFT-NUM(TF)	**0.772**	DFF-NUM(TF)	**0.706*****
	0.922		DFT-BIN50(TF)	**0.772**	DFF-BIN50(TF)	**0.731*****
h=3	HMN	Factors	OS-DFM(F)	*0.838*	BS-DFM(F)	*0.837*
	1.014		OS-RT(F)	1.397	BS-RF(F)	0.949
	AR4		DFT-NUM(F)	**0.801****	DFF-NUM(F)	**0.761*****
	0.970		DFT-BIN50(F)	**0.818***	DFF-BIN50(F)	**0.780*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.843*	BS-DFM(TF)	*0.843*
	0.975		OS-RT(TF)	1.393	BS-RF(TF)	0.948
	CADL		DFT-NUM(TF)	**0.808****	DFF-NUM(TF)	**0.752*****
	0.943		DFT-BIN50(TF)	**0.826***	DFF-BIN50(TF)	**0.776*****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). The RTs, RFs, DFMs, DFTs and DFFs were estimated with 10 factors. For more details, see notes in Table 1.

Table A.11: Robustness Results: Ten Factors

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
	AR4		DFT-NUM(F)	**0.840**	DFF-NUM(F)	**0.828****
	0.924		DFT-BIN50(F)	**0.840**	DFF-BIN50(F)	**0.834***
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
	CADL		DFT-NUM(TF)	**0.798**$^{+}$	DFF-NUM(TF)	**0.789*****
	0.922		DFT-BIN50(TF)	**0.798**$^{+}$	DFF-BIN50(TF)	**0.794*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
	AR4		DFT-NUM(F)	**0.913**	DFF-NUM(F)	**0.908**
	0.970		DFT-BIN50(F)	**0.913**	DFF-BIN50(F)	0.910
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
	CADL		DFT-NUM(TF)	**0.883***	DFF-NUM(TF)	**0.886****
	0.943		DFT-BIN50(TF)	**0.883***	DFF-BIN50(TF)	**0.883****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; SPF-REC1-IDX partitioning variable; BIC-based pruning). For more details, see notes in Table 1.

Table A.12: Robustness Results: 1st SPF-Recession Probability Index

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
	AR4		DFT-NUM(F)	**0.840**	DFF-NUM(F)	**0.785*****
	0.924		DFT-BIN50(F)	**0.840**	DFF-BIN50(F)	**0.825*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
	CADL		DFT-NUM(TF)	**0.792*****	DFF-NUM(TF)	**0.752*****
	0.922		DFT-BIN50(TF)	**0.798**$^{+}$	DFF-BIN50(TF)	**0.787*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
	AR4		DFT-NUM(F)	**0.911**	DFF-NUM(F)	**0.870****
	0.970		DFT-BIN50(F)	**0.913**	DFF-BIN50(F)	**0.907**
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
	CADL		DFT-NUM(TF)	**0.874*****	DFF-NUM(TF)	**0.875****
	0.943		DFT-BIN50(TF)	**0.883***	DFF-BIN50(TF)	**0.880****

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; SPF-REC2-IDX partitioning variable; BIC-based pruning). For more details, see notes in Table 1.

Table A.13: Robustness Results: 2nd SPF-Recession Probability Index

	Benchmarks		**Dynamic Factor Trees**		**Dynamic Factor Forests**	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
	AR4		DFT-NUM(F)	**0.811*****	DFF-NUM(F)	**0.801*****
	0.924		DFT-BIN50(F)	**0.816****	DFF-BIN50(F)	**0.805*****
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
	CADL		DFT-NUM(TF)	**0.779****	DFF-NUM(TF)	**0.796***
	0.922		DFT-BIN50(TF)	**0.794***	DFF-BIN50(TF)	**0.777*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.910*
	1.014		OS-RT(F)	1.473	BS-RF(F)	0.954
	AR4		DFT-NUM(F)	0.924	DFF-NUM(F)	0.947
	0.970		DFT-BIN50(F)	**0.912**	DFF-BIN50(F)	**0.906***
	ARL	Trgt. Fact.	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.399	BS-RF(TF)	0.951
	CADL		DFT-NUM(TF)	1.002	DFF-NUM(TF)	0.962
	0.943		DFT-BIN50(TF)	0.897	DFF-BIN50(TF)	0.896

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; ERI-IDX partitioning variable; BIC-based pruning). Assuming the absence of a publication lag of the ERI-IDX is relaxed and missing values are inferred from a regime-switching model (see Section 4.3.8). For more details, see notes in Table 1.

Table A.14: Robustness Results: Publication Lag

	Benchmarks		Dynamic Factor Trees		Dynamic Factor Forests	
Horizon	Results		Models	rRMSFE	Models	rRMSFE
h=1	HMN	Factors	OS-DFM(F)	*0.841*	BS-DFM(F)	*0.839*
	1.001		OS-RT(F)	1.210	BS-RF(F)	0.857
			DFT-NUM-PRP[BRT]-PLUS-T-LASSO(F)	**0.838**	DFF-NUM-PRP[BRT]-PLUS-T-LASSO(F)	**0.825****
	AR4		DFT-NUM-PRP[RF]-PLUS-T-LASSO(F)	0.882	DFF-NUM-PRP[RF]-PLUS-T-LASSO(F)	**0.801*****
	0.924		DFT-NUM-PRP[LAS]-PLUS-T-LASSO(F)	0.859	DFF-NUM-PRP[LAS]-PLUS-T-LASSO(F)	**0.770*****
	ARL	Trgt. Factors	OS-DFM(TF)	*0.802*	BS-DFM(TF)	*0.800*
	0.894		OS-RT(TF)	1.325	BS-RF(TF)	0.851
			DFT-NUM-PRP[BRT]-PLUS-T-LASSO(TF)	**0.789*****	DFF-NUM-PRP[BRT]-PLUS-T-LASSO(TF)	**0.782****
	CADL		DFT-NUM-PRP[RF]-PLUS-T-LASSO(TF)	0.845	DFF-NUM-PRP[RF]-PLUS-T-LASSO(TF)	**0.771*****
	0.922		DFT-NUM-PRP[LAS]-PLUS-T-LASSO(TF)	0.858	DFF-NUM-PRP[LAS]-PLUS-T-LASSO(TF)	**0.756*****
h=3	HMN	Factors	OS-DFM(F)	*0.914*	BS-DFM(F)	*0.909*
	1.014		OS-RT(F)	1.474	BS-RF(F)	0.953
			DFT-NUM-PRP[BRT]-PLUS-T-LASSO(F)	0.948	DFF-NUM-PRP[BRT]-PLUS-T-LASSO(F)	0.916
	AR4		DFT-NUM-PRP[RF]-PLUS-T-LASSO(F)	1.147	DFF-NUM-PRP[RF]-PLUS-T-LASSO(F)	0.942
	0.970		DFT-NUM-PRP[LAS]-PLUS-T-LASSO(F)	0.975	DFF-NUM-PRP[LAS]-PLUS-T-LASSO(F)	0.930
	ARL	Trgt. Factors	OS-DFM(TF)	*0.894*	BS-DFM(TF)	*0.892*
	0.975		OS-RT(TF)	1.400	BS-RF(TF)	0.951
			DFT-NUM-PRP[BRT]-PLUS-T-LASSO(TF)	0.905	DFF-NUM-PRP[BRT]-PLUS-T-LASSO(TF)	0.901
	CADL		DFT-NUM-PRP[RF]-PLUS-T-LASSO(TF)	1.011	DFF-NUM-PRP[RF]-PLUS-T-LASSO(TF)	0.906
	0.943		DFT-NUM-PRP[LAS]-PLUS-T-LASSO(TF)	0.911	DFF-NUM-PRP[LAS]-PLUS-T-LASSO(TF)	0.910

Notes: The table entries show the relative RMSFE of a particular model against the AR2 benchmark (settings: recursive scheme; DCO interpolation; Jan. 1998 first vintage; PRP-IDX partitioning variable; no pruning; DFFs use 200 bootstrapped samples). The acronym PRP indicates that the partitioning variable is based on predicted recession probabilities (two-stage machine learning framework), which are themselves derived either via boosted regression trees [BRT], random forest [RF] and lasso [LAS] algorithms and subsequently smoothened. The acronym PLUS-T indicates that time is incorporated in addition to the PRP-index. DFTs and DFFs apply Lasso-based regularization. For more details, see notes in Table 1.

Table A.15: Extended Results - Two-Stage Machine Learning Framework: PRP-based DFTs and DFFs with Lasso-based Penalization and Generalized Partitioning Sets (relative RMSFE)