Local Predictability in High Dimensions*

Philipp Adämmer[1], Sven Lehmann[2], and Rainer A. Schüssler†[3]

[1] *University of Greifswald*

[2] *University of Rostock*

[3] *University of Rostock*

July 1, 2023

Abstract

We propose a novel time series forecasting method designed to handle vast sets of predictive signals, many of which are irrelevant or short-lived. The method transforms heterogeneous scalar-valued signals into candidate density forecasts via time-varying coefficient models, and subsequently, combines them into a final density forecast via time-varying subset combination. Our approach is computationally fast, because it uses online prediction and updating. We validate our method through simulation analyses and apply it to forecast daily aggregate stock returns as well as quarterly inflation, using over 12,000 and over 400 signals, respectively. We find superior forecasting performance and lower computation time for our approach compared to competitive benchmark methods.

Keywords: Big data; Ensemble learning; Time series; Stock returns; Inflation

*We appreciate the valuable comments from participants of the 5th Quantitative Finance and Financial Econometrics Conference in Marseille. We gratefully acknowledge financial support from the German Research Foundation (DFG, 468715873).

†Corresponding author. Send correspondence to Rainer A. Schüssler. University of Rostock, Faculty of Economics and Social Sciences; e-mail: rainer.schuessler@uni-rostock.de; phone: +49 381 498-4316.

1 Introduction

Handling predictive signals that are locally concentrated in time poses a major challenge for time series forecasting models in both macroeconomics and finance (see, e.g., Stock and Watson, 2003; Giacomini and Rossi, 2010; Borup et al., 2023; Farmer et al., 2023). This applies particularly in the age of big data, in which researchers have myriads of signals at their disposal. In this paper, we introduce a forecasting method that can handle both high-dimensional signals and local predictability. Our method does not rely on a preconceived notion regarding the expected duration of each predictive signal, enabling to identify signals irrespective of their potential brevity or longevity. As our forecasting method first generates a suite of different density forecasts, and subsequently combines them into an aggregate density forecast, it can be classified as an ensemble learning approach.[1]

Although off-the-shelf machine learning methods such as Random Forests (Breiman, 2001) or eXtreme Gradient Boosting (Chen and Guestrin, 2016) can process high-dimensional predictive indicators, they are not designed for short-lived signals. The dynamic variable selection approach of Koop and Korobilis (2023) and the time-varying (point) forecast combination approach of Chen and Maung (2023) represent different strategies for dealing with short-lived predictability in high-dimensions. The latter approaches were applied to over 400 and 100 indicators, respectively. When scaling up to higher dimensions, however, the approaches of Koop and Korobilis (2023) and Chen and Maung (2023) encounter computational bottlenecks. Our ensemble learning method, on the other hand, can accommodate tens of thousands of predictive signals, integrating both information and forecasts at the same time.[2]

As a methodological innovation, our forecasting method can map vast numbers of scalar-valued signals of any type into an aggregate density forecast in a time-varying and computationally fast manner. The method proceeds in two steps. First, it maps predictive

[1] We implemented our approach in an `R`-package, using `Rcpp` (Eddelbuettel and François, 2011), `RcppArmadillo` (Eddelbuettel and Sanderson, 2014), and parallelization for efficiency. The package is publicly available on `CRAN` and and `GitHub`.

[2] See Huang and Lee (2010) for analytical, simulation-based, and empirical comparisons of both approaches.

signals into univariate density forecasts via time-varying coefficient models, where each model generates, for each signal, a conditionally normal predictive density at each point in time. The time-varying coefficient models can accommodate predictive signals of any type, including fluctuations in oil prices, text-based quantities, extracted factors, as well as lagged values and/or point forecasts of the target variable itself. Point forecasts, in turn, can be based on surveys, theory-driven models, and statistical or machine learning models. In cases where the predictive signal is a point forecast, bias can be corrected via a time-varying intercept.

In the second step, for each period, our approach selects a subset of candidate density forecasts based on their past predictive likelihoods, emphasizing recent performance by exponentially discounting predictive likelihoods (Raftery et al., 2010; Koop and Korobilis, 2012; Del Negro et al., 2016; Beckmann et al., 2020; Bernaciak and Griffin, 2022). All candidate densities within the selected subset are equally weighted. Any candidate forecast that is given zero weight at one point in time can re-enter the subset at another point in time, a crucial feature for capturing short-lived signals. The size of the subset, the choice of the candidate forecasts, and the exponential discount factor for down-weighting past performance are all data-driven. The ability of our method to accommodate candidate density forecasts based on heterogeneous signals facilitates possible diversification gains from combination (see, e.g., Timmermann, 2006; Grushka-Cockayne et al., 2017; Atiya, 2020; Kang et al., 2022). As our combination scheme allows for fast adaptation, but avoids estimating combination weights by equally weighting the density forecasts within the subsets, we follow the advice of Zellner et al. (2002) and Wang et al. (2022) to keep the combinations "sophisticatedly simple".[3]

As an alternative to exponential discounting, other time-varying weighting schemes have been proposed. For point forecast combination, these include parametric regres-

[3]Although many sophisticated weighting schemes for combining forecasts have been proposed, the simple average of forecasts often exhibits empirically superior out-of-sample performance in finite samples−the so-called "forecast combination puzzle" (Stock and Watson, 2004; Smith and Wallis, 2009; Claeskens et al., 2016; Chan and Pauwels, 2018). The key issue here is the estimation error of the combination weights that plagues sophisticated weighting schemes, whereas simple averages avoid any weight estimation. While the forecast combination puzzle has typically been studied in settings where point forecasts are combined, similar results have been found for density forecasts (see, e.g., Amisano and Geweke, 2017).

sions with regime-switching and smooth transitions (see, e.g., Deutsch et al., 1994; Elliott and Timmermann, 2005), as well as nonparametric kernel regressions (Chen and Maung, 2023). Such regression-based approaches, however, involve numerical optimization that is computationally demanding and represents a bottleneck for large sets of candidates. Similarly, existing time-varying combination schemes for density forecasts, such as those of Billio et al. (2013); Del Negro et al. (2016); McAlinn and West (2019) require simulation-based inference, which becomes a bottleneck when dealing with large sets of candidates. In contrast, our forecasting method is computationally fast, because in addition to being trivially parallelizable, online prediction and updating are feasible in both steps by using exponential discounting. We thereby avoid any expensive computations such as numerical optimization, large matrix inversion, or simulation-based inference. Another advantage of our method is that it requires minimal user input, mainly regarding the selection of grids for the tuning parameter values, which may be selected in a fully data-driven manner, or on the basis of domain-specific knowledge or recommendations from previous studies.

Although sophisticated forecasting methods should (asymptotically) beat simple forecasting methods, they often struggle to do so in finite samples.[4] To get the best of both worlds, our forecasting method accommodates complex dynamics when empirically needed, but collapses (temporarily) to simple dynamics when complexity is not required. We let the data itself reveal how many and which predictive signals are useful at each point in time, without having to decide a priori whether the predictive relationship between a given signal and the target variable is constant, evolves gradually, or changes abruptly.

We conduct simulation analyses and forecast (i) daily aggregate stock returns and (ii) quarterly inflation rates. For daily stock returns, we extend the study of Farmer et al. (2023) to high dimensions, using over 12,000 predictive signals, most of which are extracted from textual data. We find that text-based indicators have provided valuable signals over the last two decades, a period in which predictive signals based on economic

[4] As examples of simple, yet hard-to-beat benchmarks, consider the prevailing historical mean for predicting aggregate stock returns (Welch and Goyal, 2008) or autoregressive models for predicting macroeconomic variables.

indicators have largely disappeared. To predict inflation, we use a dataset of over 400 predictors compiled by Koop and Korobilis (2023) from various data sources. We also include point forecasts of inflation as predictive signals generated by the dynamic variable selection approach of Koop and Korobilis (2023), and by other methods, such as Gaussian process regressions. In the simulations and both applications, our forecasting approach is, overall, more accurate and faster than competitive benchmark methods.

The remainder of the paper is organized as follows. Section 2 lays out our methodology. Section 3 demonstrates our forecasting method in a simulation study. Section 4 presents and discusses our applications, and Section 5 concludes. Additional analyses and robustness checks are relegated to an Appendix.

2 Methodology

In this section, we first outline the structure of the candidate forecasting models and then present our proposed subset combination.

2.1 Candidate density forecasts

We generate each candidate density forecast based on a univariate time-varying coefficient (TV-C) model that can be written in state-space form:

$$y_t = z_t\theta_t + \varepsilon_t, \quad \varepsilon_t \overset{ind}{\sim} \mathcal{N}(0, H_t) \tag{1}$$

$$\theta_t = \theta_{t-1} + \zeta_t, \quad \zeta_t \overset{ind}{\sim} \mathcal{N}(0, W_t)\,. \tag{2}$$

Equation (1) is the observation equation, and Equation (2) is the system (state) equation. The target series is y, and $z_t = [1, s_{t-1}]$ denotes a vector which includes an intercept and one predictive signal s.[5] The predictive signal can be of any type, ranging from simple predictors such as exchange rate or oil price fluctuations, to extracted factors,

[5]We specify our methodology for one-step-ahead forecasts without loss of generality. Direct forecasts could be obtained for alternative forecast horizons $h > 1$. By including one signal in each TV-C model, signals with different lengths can be accommodated and the appropriate degree of time variation can be selected for each signal.

to point forecasts of y_t based on information through $t-1$. As the signal itself may be based on multivariate information (e.g., by extracted factors or point forecasts of y that are based on multivariate information) and/or non-linear transformations, we do not sacrifice flexibility. Let θ_t denote a 2×1 vector of coefficients (states). The errors ε_t and ζ_t are assumed to be mutually independent for all leads and lags, and the coefficients evolve according to a multivariate random walk (see, e.g., Cogley and Sargent, 2005; Dangl and Halling, 2012; Koop and Korobilis, 2012; Beckmann et al., 2020). For given values of H_t and W_t, standard Kalman filtering results can be applied to carry out recursive estimation and forecasting. To specify the time-varying observational variance H_t and system covariance matrix W_t, we use a discount factor approach (see, inter alia, West and Harrison, 1997; Raftery et al., 2010; Dangl and Halling, 2012; Koop and Korobilis, 2012; Hill and Rodrigues, 2022). Kalman filtering consists of iteratively applying a prediction and an update step. All models are estimated independently from each other, and, for ease of notation, we suppress model indices in this subsection.

Suppose that $\theta_{t-1|t-1} \sim \mathcal{N}\left(\widehat{\theta}_{t-1}, \Sigma_{t-1}\right)$. Then, the prediction step involves:

$$\theta_{t|t-1} \sim \mathcal{N}\left(\widehat{\theta}_{t-1}, R_t\right), \tag{3}$$

where

$$R_t = \Sigma_{t-1} + W_t. \tag{4}$$

Instead of estimating the system covariance matrix W_t, we use a discount factor λ that controls the dynamics of the coefficients:

$$R_t = \frac{\Sigma_{t-1}}{\lambda}. \tag{5}$$

This modeling approach involves exponential discounting for which data that is τ time points old has weight λ^τ, and the effective window size is $(1-\lambda)^{-1}$. Within each model, we fix the value of λ from an application-specific grid of possible values $\mathscr{S}_\lambda$ in our empirical work. A value of $\lambda = 1$ corresponds to constant coefficients and is a natural

upper bound. Lower values of λ are associated with more time variation in the coefficients. In our simulation and two applications, we choose the lower bound to match an effective window size of 2.5 years, which corresponds to highly volatile coefficients and is similar to the choice of lower bounds in previous studies. (see, e.g., Koop and Korobilis, 2012; Dangl and Halling, 2012). For daily data, we set $\mathscr{S}_\lambda = \{0.9984,\ 0.9992,\ 1.0000\}$, for monthly data $\mathscr{S}_\lambda = \{0.9667,\ 0.9833,\ 1.0000\}$, and for quarterly data $\mathscr{S}_\lambda = \{0.90,\ 0.95,\ 1.00\}$. Each predictive signal in combination with each value of λ defines a separate TV-C model that generates a candidate density forecast. Once we have observed y_t, we update the estimates of the coefficients and their covariance:

$$\theta_{t|t} \sim \mathcal{N}\left(\widehat{\theta}_t, \Sigma_t\right), \tag{6}$$

with

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} + \underbrace{R_t z_t' \left(H_t + z_t R_t z_t^{'}\right)^{-1}}_{\text{Kalman gain}} \underbrace{\left(y_t - z_t \widehat{\theta}_{t-1}\right)}_{\text{e}_t}, \tag{7}$$

where e_t denotes the time-t one-step ahead prediction error, $z_t R_t z_t^{'}$ refers to the variance resulting from estimation uncertainty of θ_t, and

$$\Sigma_t = R_t - \underbrace{R_t z_t' \left(H_t + z_t R_t z_t^{'}\right)^{-1}}_{\text{Kalman gain}} z_t R_t. \tag{8}$$

Forecasting can be done based on the conditionally normal predictive density:

$$y_{t|t-1} \sim \mathcal{N}\left(\underbrace{z_t \widehat{\theta}_{t-1}}_{\mu_t}, \underbrace{H_t + z_t R_t z_t^{'}}_{\sigma_t^2}\right). \tag{9}$$

To accommodate time-varying volatility, a stylized fact of financial and macroeconomic time series, we use an Exponentially Weighted Moving Average (EWMA) estimate of the observational variance H_t:

$$\widehat{H}_t = (1-\kappa) \sum_{\tau=1}^{t-1} \kappa^{\tau-1} \left(y_{t-\tau} - z_{t-\tau} \widehat{\theta}_{t-\tau}\right)^2. \tag{10}$$

We can recursively approximate the EWMA specification in Equation (10) to obtain volatility forecasts:

$$\widehat{H}_{t+1|t} = \kappa \widehat{H}_{t|t-1} + (1-\kappa)(y_t - z_t\widehat{\theta}_t)^2. \tag{11}$$

Following the recommendation of RiskMetrics (Reuters, 1996), we set the discount factor to $\kappa = 0.94$ for daily data, $\kappa = 0.97$ for monthly data, and to $\kappa = 0.98$ for quarterly data in our applications. We set $\widehat{H}_1 = \widehat{Var}\left(y_{initial}\right)$, where $\widehat{Var}\left(y_{initial}\right)$ denotes the estimate of the variance of y, computed over an initial sample of five years. Note from Equation (7) that the Kalman gain determines how the model learns from past forecast errors: the higher the value of the Kalman gain, the higher the adaptiveness to new data. Although stochastic volatility or (G)ARCH specifications for modeling the dynamics of H_t are more sophisticated choices, they would increase the computational burden substantially, since Markov chain Monte Carlo methods or numerical optimization would be needed.[6] The volatility dynamics produced by the simple EWMA are generally similar to those of more sophisticated alternatives and, empirically, differences between other specifications of the volatility dynamics are usually not substantial in macroeconomic and financial time series applications (see, e.g., Koop and Korobilis, 2012; Clark and Ravazzolo, 2015; Cederburg et al., 2023). Note that conditionally normal predictive densities with time-varying volatility are compatible with unconditionally leptokurtic observations, a stylized fact of financial time series.

We initialize $\widehat{\theta}_0 = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma_0 = \begin{pmatrix} v_1^2 & 0 \\ 0 & v_2^2 \end{pmatrix}$. In addition, we follow Raftery et al. (2010) and set $\mu_1 = \mu_2 = 0$, $v_1^2 = \widehat{\beta}_0^2 + \widehat{Var}\left(y_{initial}\right)$ and $v_2^2 = \widehat{Var}\left(y_{initial}\right) / \widehat{Var}\left(x_{initial}\right)$, where $\widehat{\beta}_0$ denotes the estimated intercept of a fitted static linear regression over an initial sample of five years. If s represents a point forecast of y, we fix the slope coefficient to one in our applications, but allow for a time-varying intercept, so as to accommodate (possibly time-varying) biases of the point forecast by setting $\mu_1 = 0$, $\mu_2 = 1$, $v_1^2 = \widehat{\beta}_0^2 + \widehat{Var}\left(y_{initial}\right)$

[6] Another alternative is to model the observational variance as an inverse-gamma distribution, which would result in a conditionally t-distributed target variable. However, this would prohibit logarithmic combination, because the t-distribution is not a member of the exponential family. We outline the advantages of the logarithmic combination rule in 2.2 and why we use it in our approach.

and $v_2^2 = 0$ in our applications. If researchers, however, assume that the point forecast is unbiased, they will set $v_1^2 = 0$.

2.2 Time-varying subset combination

Let $\mathscr{M}$ denote the set of candidate density forecasts, which we take as given in the combination step. We index the TV-C models at our disposal as $M_j, j = 1, \ldots, J$. To predict the next period's outcome, we select a subset of TV-C models that have generated high predictive likelihoods relative to other TV-C models in the (recent) past. Each period, we rank the TV-C models according to their generated sum of discounted predictive log-likelihoods (DPLLs) until the given point in time. For a given value of the discount factor γ, which exponentially down-weights past predictive log-likelihoods, we compute the DPLL of the j-th TV-C model as:

$$DPLL_{j,t|t-1}(\gamma) = \sum_{\tau=1}^{t-1} \gamma^{\tau} \cdot \ln\left[p_j\left(y_{t-\tau|t-\tau-1}\right)\right], \tag{12}$$

where $p_j\left(y_{t-\tau|t-\tau-1}\right)$ denotes its predictive likelihood at time $t-\tau$. We then assign equal weights (that sum to one) to an active subset of the TV-C models with the highest DPLLs. The tuning parameter ψ controls the size of the active subset. Let $rk(1), rk(2), \ldots, rk(\psi)$ be the TV-C model with the highest, second highest and ψ-th highest DPLL in the active subset, respectively. The weights of the candidate density forecasts that are not (temporarily) part of the active subset equal zero.

Let $\mathscr{M}_t^* \subseteq \mathscr{M}$ denote the time-t active subset of TV-C models. We index the time-t active TV-C models as $M_{i,t}^*, i = 1, \ldots, \psi_t^*$, where ψ_t^* denotes the optimized time-t subset size. Each period we choose the combination of γ and ψ that would have maximized the sum of discounted predictive log-likelihoods of the *combined* predictive densities until $t-1$:

$$\left(\gamma_{t|t-1}^*, \psi_{t|t-1}^*\right) = \underset{(\gamma_t,\psi_t)|\gamma_t \in S_\gamma, \psi_t \in S_\psi}{\arg\max} \sum_{\tau=1}^{t-1} \delta^{\tau} \cdot \ln\left[p_{comb}^{(\gamma_t,\psi_t)}\left(y_{t-\tau|t-\tau-1}\right)\right], \tag{13}$$

where $p_{comb}^{(\gamma_t,\psi_t)}\left(y_{t-\tau|t-\tau-1}\right)$ denotes the predictive likelihood of the combined predictive

density (using logarithmic combination) at time $t-\tau$, which depends on the choice of time-dependent tuning parameters γ_t and ψ_t. We choose the combination of γ_t and ψ_t in (13) from the two-dimensional grid $\mathscr{S}_\gamma \times \mathscr{S}_\psi := \{(\gamma_t, \psi_t) \,|\gamma_t \in S_\gamma, \psi_t \in S_\psi\}$ in a time-dependent manner. In our empirical work, we select from a broad range of possible values in $\mathscr{S}_\gamma$, covering from rapid model switching to a recursive weighting scheme ($\gamma = 1$). In our applications and the simulation study, we set $\mathscr{S}_\gamma = \{0.40{:}0.10{:}0.90, 0.90{:}0.01{:}1.00\}$. Similarly, we use a broad range of values for $\mathscr{S}_\psi$, covering from pure model selection ($\psi = 1$) to a combination of many or even all candidate density forecasts. In the simulation study and application to quarterly inflation forecasting, we choose $\mathscr{S}_\psi = \{1{:}1{:}100\}$. In the application to daily forecasting, we set $\mathscr{S}_\psi = \{1{:}1{:}10, 10{:}10{:}100\}$, adopting a coarser grid due to the vast number of predictive signals.

We use logarithmic combination, because the combined predictive distribution retains the same distribution of its components, whereas linear combination does not (see, e.g., Faria and Mubwandarikwa, 2008).[7] We emphasize the recent forecasting performance when optimizing the values of γ and ψ by using a discount factor δ. In our empirical work, we choose the value of δ such that the effective window size is five years, that is, we set $\delta = 0.9992$ in case of daily data frequency, $\delta = 0.9833$ in case of monthly data frequency, and $\delta = 0.95$ in case of quarterly data frequency. Yet we find our results largely unchanged for alternative choices of δ.

Using the logarithmic combination rule, we obtain a normally distributed time-t density forecast as:

$$y_{t|t-1} \sim \mathcal{N}\left(\mu_{t,comb}, \sigma^2_{t,comb}\right), \tag{14}$$

[7] The candidate density forecasts are of similar shape in our setup, because each of them is specified with the same value of the tuning parameter κ, which controls the dynamics of the observational variance H_t in Equation (10). The observational variance, in turn, makes up the lion's share of the conditional variance σ^2_t. Differences in the conditional means across the candidate forecasting models are dominated by the conditional variances, which are similar across the candidates. Hence, the shapes of the candidate predictive densities are similar and the mixture density based on a linear combination will be of similar shape as well. The mixture density obtained from a linear combination would thus look very similar to a normal distribution in our setting, but would be more cumbersome to evaluate for large sets of candidate predictive densities than using the logarithmic combination. See West and Harrison (1997), p. 438, for an illustration of a mixture density that is similar to the shape of its components. Furthermore, if the candidate density forecasts are appropriately calibrated, the linearly combined density forecast exhibits excessive dispersion (see, e.g., Smith and Wallis, 2009). The logarithmic combination rule, in contrast, does not have this drawback.

with

$$\frac{\mu_{t,comb}}{\sigma^2_{t,comb}} = (\psi^*_t)^{-1} \sum_{k=1}^{\psi^*_t} \frac{\mu_{rk(k),t}}{\sigma^2_{rk(k),t}} \tag{15}$$

and

$$\sigma^{-2}_{t,comb} = (\psi^*_t)^{-1} \sum_{k=1}^{\psi^*_t} \sigma^{-2}_{rk(k),t}, \tag{16}$$

where the mean $\mu_{rk(k),t}$ and variance $\sigma^2_{rk(k),t}$ of the k-th-ranked TV-C model are in each case computed from information until $t-1$. The density forecasts are strictly (pseudo) out-of-sample (OOS), since we use only information that would have been available at a given point in time.

Our time-varying subset combination can be seen as an alternative to Markov switching on the model space, but is less prone to estimation error and computationally much more convenient, since online updating and prediction is computationally feasible without having to specify a transition matrix and re-estimate the model from scratch in each period.[8] We do not implicitly assume that one of the candidate forecasting models represents the true data-generating process, since the size and composition of the chosen subset may change over time and does not necessarily converge to one specific candidate, even asymptotically. This differs, for example, from Bayesian model averaging, where the combination asymptotically collapses to a particular model whose predictive probabilities dominate all others, which can lead to undesired outcomes in the (realistic) case of a misspecified model set (Diebold, 1991).

In our applications, we do not exclude any predictive signal (and its associated forecast) *a priori* based on some pre-selection step (e.g., Lasso-type regressions), because it may well be that a given predictive signal has no predictive power over a training sample, but may provide useful signals locally over time. Our time-varying subset selection allows for local predictability, since candidate predictive densities can be switched off at particular points in time, but subsequently switched on again.

It is instructive to present our time-varying subset selection from the perspective of a constrained optimization problem to see more clearly the role of the ranking procedure.

[8] See Raftery et al. (2010) for a more in-depth connection between Markov switching methods on the model space and parsimoniously parameterized alternatives for model changes.

The optimization problem can be stated as follows:

$$\mathbf{w}_t^* = \underset{\mathbf{w_t} \in \mathcal{W}}{\arg\max} \sum_{\tau=1}^{t-1} \gamma^\tau \cdot \ln \left[p_{comb}^{(\gamma_t, \psi_t)} \left(y_{t-\tau|t-\tau-1} \right) \right] \tag{17}$$

$$\text{s.t.} \quad \sum_{j=1}^{J} \mathbb{1}_{(w_{j,t} \neq 0)} = \psi_t, \tag{18}$$

$$w_{j,t} \in \left\{ 0, \psi_t^{-1} \right\}, \forall j, \tag{19}$$

where $\mathbf{w}_t = (w_{1,t}, \ldots, w_{J,t})'$ are the combination weights and $\mathbb{1}_{(\bullet)}$ denotes the indicator function. The objective (17) with the constraints (18) and (19) is a best subset optimization problem, in which all candidate forecasts within the subset are equally weighted and sum to one. The binary optimization problem falls into the class of computationally tedious NP-hard problems. As we work with time series data and have to elicit tuning parameters, we have to solve this optimization problem at each point in time and for different combinations of values for γ and ψ. As a remedy, we replace the constraints (18) and (19) by a ranking procedure and assign the weights:

$$w_{rk(1),t}, \ldots, w_{rk(\psi_t),t} = \psi_t^{-1}, \tag{20}$$

setting the remaining weights to zero. We thereby avoid any optimization when selecting the subset for a given combination of of γ and ψ.

Our time-varying subset procedure builds on the findings of Diebold and Shin (2019), who provide an interesting perspective on best subset averaging. The authors propose "partially-egalitarian Lasso" for forecast combination, for which, in a first step, they discard a fraction of the candidate forecasts based on Lasso regressions, and in a second step, shrink the survivors toward equal weights. Shrinking the survivors to exactly equal weights is found to be superior. Similarly, when exploring regularized mixtures of predictive densities, Diebold et al. (2022) find simple averaging of subsets to be empirically successful and that there is little gain from regularization beyond best subset averaging.

Regarding the selection step, Diebold and Shin (2019) find no empirical advantage of

a portfolio perspective over an individual ranking procedure, despite working with only a small number of candidate forecasts. The portfolio perspective uses the best-performing average that takes into account the dependency structure between the candidate forecasts, which corresponds to the optimization problem stated in (17) to (19). Hence, replacing the constraints (18) and (19) with constraint (20) appears empirically reasonable, especially for large sets of candidate forecasts, which would otherwise result in high computational burdens.

Although the candidate forecasts that are selected in our ranking procedure are based on their individual performance, dependencies between the candidates' forecasting performance are implicitly taken into account, because we elicit the subset size ψ in (13) in a data-driven and time-varying manner. Hence, the selected subset size in a given period reflects the empirical performance based on the portfolio of selected candidate density forecasts. The key here is that selection and weighting are done jointly in *one* step for all combinations of ψ and γ from the two-dimensional grid $\mathscr{S}_\gamma \times \mathscr{S}_\psi$, whereas in the literature, a two-step procedure prevails. The latter uses elimination procedures as a pre-screening step before combining the remaining forecasts, where the selection step is detached from the subsequent weighting step (see, e.g., Aiolfi and Favero, 2005; Samuels and Sekkel, 2017; Diebold and Shin, 2019; Chen and Maung, 2023).[9] In sum, the second part of our approach extends best subset averaging to i.) a one-step approach with the subset size being based on a cross-validated tuning parameter, and ii.) time-varying weights by exponentially discounting past performance which is based on cross-validated tuning parameters. Finally, we apply the combination procedure to density rather than to point forecasts.

We call our forecasting method "signal-transformed subset combination", hereafter STSC, and compare it with four state-of-the-art machine learning methods in the simulation study and our two applications: Random Forests (Breiman, 2001), hereafter RF, Boosted Regression Trees (Friedman, 2001, 2002), hereafter BRT, eXtreme Gradient Boosting (Chen and Guestrin, 2016), hereafter XGB, and Relaxed Lasso (Meinshausen,

[9] A notable exception is Roccazzella et al. (2022), who combine (point) forecasts by using constrained optimization with penalty in a one-step procedure.

2007), hereafter RLasso. We choose these benchmarks because they are computationally feasible for vast numbers of predictive signals, and because of their documented strong performance in predicting macroeconomic and financial time series in the presence of high-dimensional predictors. (see, e.g., Gu et al., 2020; Hastie et al., 2020; Bianchi et al., 2021; Medeiros et al., 2021).[10]

3 Simulation study

We conduct a simulation study to compare the predictive accuracy and computational speed of our forecasting method with competitive benchmark methods. To do so, we consider six data-generating processes (DGPs), generated from TV-C models that differ with respect to the predictive relationship between y and signal s, spanning constant, gradually evolving, and abruptly changing relationships. The signals at our disposal are indexed as $i = 1, \ldots, k$. The stochastic processes are generated as follows:

$$y_t = \sum_{i=1}^{k} \theta_{i,t} s_{i,t-1} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right),$$

Evolution of the coefficients in DGP 1:

$$\theta_{1,t} = \begin{cases} -0.5, & t > 0. \end{cases}$$

$$\theta_{1:k,t} \backslash \theta_{1,t} = \begin{cases} 0, & t > 0. \end{cases}$$

[10] We implemented RF, BRT and XGB in `R` using `LightGBM` and `xgboost`, respectively. RLasso was implemented using the coordinate descent algorithm from Friedman et al. (2010), implemented in the `R` package `glmnet`. For all methods, we used default choices for the tuning parameters.

Evolution of the coefficients in DGP 2:

$$\theta_{2,t} = \begin{cases} -0.4, & 200 < t < 450 \\ 0.5, & \text{otherwise.} \end{cases}$$

$$\theta_{1:k,t} \backslash \theta_{2,t} = \begin{cases} 0, & t > 0. \end{cases}$$

Evolution of the coefficients in DGP 3:

$$\theta_{3,t} = \begin{cases} 0.8 - 0.5 \cdot t/420, & t < 420 \\ -0.2 + 0.5 \cdot t/420, & \text{otherwise.} \end{cases}$$

$$\theta_{1:k,t} \backslash \theta_{3,t} = \begin{cases} 0, & t > 0. \end{cases}$$

Evolution of the coefficients in DGP 4:

$$\theta_{4,t} = \begin{cases} 0, & t < 400 \\ 0.50, & \text{otherwise.} \end{cases}$$

$$\theta_{1:k,t} \backslash \theta_{4,t} = \begin{cases} 0, & t > 0. \end{cases}$$

Evolution of the coefficients in DGP 5:

$$\theta_{5,t} = \begin{cases} -2 \times 10^{-3} \cdot t, & t < 400 \\ 0, & \text{otherwise.} \end{cases}$$

$$\theta_{1:k,t} \backslash \theta_{5,t} = \begin{cases} 0, & t > 0. \end{cases}$$

Evolution of the coefficients in DGP 6:

$$\theta_{6,t} = \begin{cases} 0.6 - 0.2 \cdot t/180, & 60 < t < 180 \\ 0.6 + 0.2 \cdot t/420, & 250 < t < 420 \\ 0, & \text{otherwise.} \end{cases}$$

$$\theta_{1:k,t} \backslash \theta_{6,t} = \begin{cases} 0, & t > 0. \end{cases}$$

We simulated $t = 1, \ldots, 500$ data points and set the number of available signals k to 501. Furthermore, we investigated the (relative) performance of our method for different noise levels $\sigma_\varepsilon^2 \in \{0.1, 0.5, 1.0\}$. Given the number of data points, we treat the data as "monthly" and choose the (grids of) tuning parameters for STSC accordingly.

First, we benchmark our method with respect to predictive accuracy. To do so, data points $t = 1, \ldots, 49$ were used for training and $t = 50, \ldots, 500$ were used for OOS predictions. We ran the simulation 100 times and computed the average mean squared errors (MSEs) across all runs for each method. Table 1 summarizes the results. The MSE of each method is divided by the MSE of STSC. A value greater than one therefore indicates a more accurate prediction by STSC. With few exceptions, STSC prevails consistently, doing comparatively well at capturing different (time-varying) relationships, and in discarding signals that are pure noise.

Table 1: Relative forecast accuracy results.

	BRT	STSC	RF	RLasso	XGB
$\sigma_\varepsilon^2 = 0.1$					
DGP 1	1.28	1.00	1.26	**0.99**	1.35
DGP 2	1.82	**1.00**	1.99	1.92	2.13
DGP 3	1.37	**1.00**	1.54	1.18	1.51
DGP 4	1.29	**1.00**	1.29	1.25	1.43
DGP 5	1.60	**1.00**	1.52	1.42	1.67
DGP 6	1.69	**1.00**	1.60	1.45	1.78
$\sigma_\varepsilon^2 = 0.5$					
DGP 1	1.12	1.00	1.09	**0.98**	1.24
DGP 2	1.17	**1.00**	1.23	1.19	1.36
DGP 3	1.12	**1.00**	1.13	1.02	1.26
DGP 4	1.03	**1.00**	1.06	1.03	1.20
DGP 5	1.19	**1.00**	1.17	1.11	1.32
DGP 6	1.19	**1.00**	1.17	1.09	1.32
$\sigma_\varepsilon^2 = 1.0$					
DGP 1	1.08	1.00	1.06	**0.98**	1.20
DGP 2	1.06	**1.00**	1.10	1.07	1.21
DGP 3	1.07	1.00	1.08	**0.99**	1.21
DGP 4	1.00	**1.00**	1.03	1.00	1.14
DGP 5	1.08	**1.00**	1.08	1.05	1.21
DGP 6	1.10	**1.00**	1.09	1.04	1.22

The entries correspond to each model's MSE divided by the MSE of STSC. The MSEs are computed across 100 simulation runs, each with 451 OOS observations.

Next, we benchmark STSC in terms of computation time. Table 2 summarizes the relative computation time of each method relative to STSC for different numbers of observations (n) and signals (p). We show the (relative) computation time for generating one and five OOS predictions, respectively. The table clearly shows the benefits of our online updating and prediction approach: while STSC is faster than all benchmarks except RLasso and RF for a single prediction, STSC's speed advantage improves vastly as more subsequent predictions are computed. Note that the benchmarks were run on a single core (Apple M1). As STSC uses parallel computing, the absolute times in real-world applications are even lower.

Table 2: Timing results.

	n	p	BRT	STSC	RLasso	RF	XGB
Single Prediction							
	500	500	1.27	1.00 (0.50)	1.35	**0.99**	3.45
	1,000	1,000	2.90	**1.00** (1.14)	1.60	2.47	8.31
	5,000	5,000	1.92	**1.00** (21.64)	**1.00**	1.81	10.56
	10,000	10,000	1.78	1.00 (93.86)	**0.84**	1.58	9.60
Five Predictions							
	500	500	6.64	**1.00** (0.45)	6.68	5.13	19.05
	1,000	1,000	14.62	**1.00** (1.10)	7.85	12.12	42.91
	5,000	5,000	9.60	**1.00** (21.53)	4.92	9.01	53.39
	10,000	10,000	9.12	**1.00** (94.00)	4.32	7.95	48.49

The table shows the (relative) time required to generate a single (or five consecutive) OOS prediction(s). Computation time is shown relative to STSC. The runtime in seconds (each time averaged over ten repetitions) for STSC is shown in parentheses. The results are based on $\sigma_\varepsilon^2 = 0.5$.

4 Empirical work

4.1 Application I: Forecasting daily aggregate stock returns

In the first application, we predict aggregate US stock returns, a topic that has been central to the field of financial economics ever since the inception of stock markets. Although numerous predictive variables have been proposed, Welch and Goyal (2008) and Goyal et al. (2023) find most of them useless for producing superior OOS forecasts compared to the prevailing historical mean. One reason for this result might be that predictors are useful for short stretches ("pockets"), but do not have predictive power most of the time (see, e.g., Paye and Timmermann, 2006; Farmer et al., 2023).

The phenomenon of short-lived predictability could arise as a consequence of time-varying risk premia or market inefficiencies. Based on kernel regressions and four economic indicators, Farmer et al. (2023) find that local pockets of predictability are consistent with sticky expectations, for which investors sluggishly update their beliefs about a persistent component in the cash flow process. Yet the authors "only find limited sup-

port" for time-varying risk premia. We extend the study of Farmer et al. (2023) to high dimensions.

4.1.1 Data

We use daily data, since local pockets are probably short-lived and, hence, more hidden at lower frequencies. The target variable is the value-weighted CRSP US stock market return minus the one-day return on a short T-bill rate. We use two sets of predictive signals: *Economic* and *Economic & Text*.

Economic

In this set, we initially follow Farmer et al. (2023), using four economic signals that are available at daily frequency: first, the lagged dividend-price ratio (*dp*), computed as dividends over the most recent 12-month period divided by the closing price of a given day. Second, the yield on a 3-month Treasury bill (*tbl*). Third, the term spread (*tsp*), computed as the difference between yields on a 10-year Treasury bond and a 3-month Treasury bill. And fourth, a variance measure (*vola*), computed as the realized variance over the previous 60 trading days.

Next, we include economic indicators as point forecasts. First, we generate a point forecast of y using a BRT based on ten volatility lags (measured by the CBOE volatility index), to capture a non-linear predictive relationship between returns and volatility, as suggested by investor flight-to-safety (see, e.g., Adrian et al. 2019).[11] We call this predictive signal *GBT*. Second, we use an equally weighted combination of the forecasts generated by *tbl* and *dp*, motivated by the findings of Tsiakas et al. (2020) that these two predictors perform well in different states of the economy. We call this signal *TBL_DP*. Third, following, inter alia, Pettenuzzo et al. (2014), we include forecasts generated by principal component regressions, for which the principal components are extracted from *dp*, *tbl*, *tsp*, and *vola*. The number of components is dynamically selected by the adjusted R^2. We call this signal *PCR*. Fourth, we include the prevailing historical mean, *PHM*, a competitive estimator of aggregate stock returns (Welch and Goyal, 2008; Goyal et al.,

[11] Before January 2, 1990, we use a realized volatility estimator as proposed by Mele (2007).

2023). In the setting *Economic*, we have $8 \times 3 = 24$ density forecasts at our disposal in each period, since the three different values of the discount factor λ in the grid $\mathscr{S}_\lambda$ define three different candidate forecasts for each signal.

Economic & Text

In addition to the signals contained in *Economic*, *Economic & Text* adds a large number of text-based indicators. Our text corpus comprises 793,013 news articles from *The New York Times* and *The Washington Post* between 1980-06-02 and 2021-12-31. The data source is the legal database LexisNexis. We have downloaded economically related newspaper articles that, for example, contain the string *econom* in the headline or body of the text.

We only retained English articles and removed stop words (e.g., *the*, *end*, *for*, etc.), punctuation, numbers and symbols. We then created a document-term matrix (dtm), where each row i corresponds to a document, and each column j to a stemmed word. Each cell indicates how often the j-th stemmed word occurred in the i-th document. We then grouped all documents per day and weighted each word count relative to all counted words per day, retaining only those rows (days) from the dtm which corresponded to a trading day. To remove very rare and misspelled words, we required each word to be included in at least 0.1% of the in-sample documents. Our final dtm consists of 10,487 rows (days) and 12,288 columns (terms), where each column serves as a signal.[12] The training sample for the textual predictors spans 1980-06-02 to 1998-12-31.[13]

Our decision to use relative word counts as a signal is mainly driven by an effort to mimic the information set of a real-time forecaster, so as to reduce data mining concerns in the spirit of Yan and Zheng (2017).[14] While more sophisticated methods such as textual factors (Cong et al., 2019), topic models (see, e.g., Thorsrud, 2020) or sentiment approaches (see, e.g., Barbaglia et al., 2022) could be used to generate text-based signals, those techniques would not have been available to a researcher at the start of our sample

[12] We used the `R`-package *quanteda* (Benoit et al., 2018) for pre-processing.

[13] We exploit the longer data histories of the economic signals by using longer training samples for them.

[14] Lima and Godeiro (2023) use (absolute) word counts for predicting stock returns, but with monthly data.

and involve a couple of subjective choices for data processing. In addition, estimating textual factors or topic proportions requires a series of design choices. As a result, word counts may provide a more objective and conservative assessment of the incremental value embedded in textual data from a real-time perspective, than more sophisticated alternatives. In the setting *Economic & Text*, we have $(8+12,288)\times 3 = 36,888$ candidate density forecasts at our disposal in each period.

4.1.2 Evaluation metrics

We evaluate the performance of STSC and the benchmark methods with both statistical and economic measures. As a measure of statistical forecast accuracy, we report Clark and West (2007) (CW) test statistics.[15]

To evaluate the economic significance of the forecasts, we follow Farmer et al. (2023) and form a managed portfolio with excess returns:

$$r^p_{t+1} = w^*_t \cdot r^m_{t+1}, \tag{21}$$

where r^m_{t+1} is the realized market excess return. The weight placed on the market is:

$$w^*_t = \left(\frac{1}{\eta}\right)\left(\frac{\widehat{r}_{t+1}}{\widehat{\sigma}^2_{t+1}}\right), \tag{22}$$

where $\widehat{r}_{t+1}$ denotes the expected excess return and $\widehat{\sigma}^2_{t+1}$ denotes its expected variance. The risk aversion parameter η is set to three. We restrict the portfolio weights between zero and two, ruling out short sales and allowing for a maximum leverage ratio of two.

We then use the excess returns obtained from the managed portfolio (21) to evaluate the risk-adjusted return α from the regression:

$$r^p_{t+1} = \alpha + \beta \cdot r^m_{t+1} + \epsilon_{t+1}. \tag{23}$$

[15]The CW test considers estimation error, which is expected to be higher in the larger model (relative to the prevailing historical mean forecast). Hence, the test summarizes the true predictive power of the indicators in the larger model.

We report the annualized estimated α in percentage points and annualized certainty equivalent returns (CERs).

4.1.3 Results

Table 3 summarizes the results for STSC and the four benchmark methods, based on the evaluation sample from 1999-01-04 to 2021-12-31.

Table 3: Application to stock returns: Forecast evaluation.

	CW	$\widehat{\alpha}$	CER
Economic			
STSC	0.14	2.54%*	5.04%
BRT	0.46	1.46%	−1.83%
RLasso	−0.15	0.00%	−5.10%
RF	0.90	3.04%	0.04%
XGB	−0.28	−4.24%	−7.93%
Economic & Text			
STSC	1.48*	4.23%**	6.08%
BRT	−0.76	−2.66%	−6.72%
RLasso	−1.09	−0.04%	−5.14%
RF	−0.49	−8.41%	−11.34%
XGB	−0.11	−0.32%	−3.77%

The table reports the Clark and West (2007) (CW) test statistics for OOS predictability measured relative to the prevailing historical mean. Further, we report the estimated annualized alpha ($\hat{\alpha}$) and the (annualized) certainty equivalent return (CER) values. One star indicates significance at the 10% level; two stars at the 5% level; and three stars at the 1% level (for one-sided alternatives). The evaluation sample spans 1999-01-04 to 2021-12-31.

For *Economic*, all methods have insignificant CW test statistics, but STSC generates, overall, higher economic values in terms of alphas and CERs compared to the benchmark methods. For *Economic & Text*, we observe a statistically significant CW test statistic at the 10% level for our approach, but not for the benchmark methods.[16] Also, in terms of economic performance, the incremental value of the text-based signals is substantial when using STSC, with considerably higher estimated alphas and CERs. In contrast, the benchmark methods cannot, overall, successfully exploit the text-based signals.

[16]In Appendix 6.2, we compare STSC to the kernel regression approach of Farmer et al. (2023) for four economic signals and an extended evaluation period starting in 1967. Similar to Farmer et al. (2023), we find that economic signals contained valuable information until the 1980s, but less so afterwards.

Figure 1 depicts the forecasting performance, the selected signals, and the subset size over time for *Economic* (left column) and *Economic & Text* (right column). The top panel shows how the cumulative sums of squared error differences (CSSEDs) between the PHM forecast and STSC have evolved over time.[17] Positive values hence indicate more accurate forecasts of STSC compared to the PHM in a mean squared error sense. While the CSSEDS are generally negative over the entire evaluation period for the setting *Economic*, we observe positive CSSEDs in the setting *Economic & Text* with an episode of substantial gains between 2002 and 2005.

The middle panel of Figure 1 depicts the signals that were selected over time. In *Economics*, all economic signals were selected at least once over time, of which *vola* was selected most often (see Figure 6 in Appendix 6.1, which provides an enlarged version of the left plot in the middle panel of Figure 1). The subset size ψ fluctuated over time (see lower left panel of Figure 1), but mostly only one candidate forecast was selected (out of 24). In *Economics & Text*, the variation in subset size was more pronounced. Nevertheless, with few exceptions, the subset size was well below the upper limit of 100 (out of 36,888 candidate forecasts), indicating a sparse structure of text-based signals. This result is further substantiated, considering that only 121 signals of many thousands were selected over the entire evaluation period.

Table 5 in the Appendix lists all 121 signals that were selected at least once over the entire evaluation sample. We observe that only text-based signals (i.e., word stems) were selected, but no economic signals. This result aligns with our extended analysis for the economic signals in Appendix 6.2, and previous studies such as Farmer et al. (2023) and Demetrescu et al. (2022) who find that the predictive power of economic signals has substantially weakened over the last few decades. Many of the text-based signals selected were part of an insurance strategy where STSC combined many signals to produce low-variance forecasts that "replicated" the PHM. These episodes hence coincide with squared errors that match those of the PHM (that is, plateaus in the CSSEDs). The "replication" of the PHM also explains why many words without any apparent connection

[17] We have omitted the CSSEDs of the benchmark methods in the CSSED plots, because they are substantially negative and hence would distort the scale.

to economics were selected. Episodes with increasing CSSEDs coincide with the selection of only a few but strong signals such as "hardwood", which can be related to the housing market. Overall, we observe different levels of persistence in the selection of text-based signals, that is, some signals are selected over longer episodes, and others only over short stretches—reflecting the flexibility of STSC.

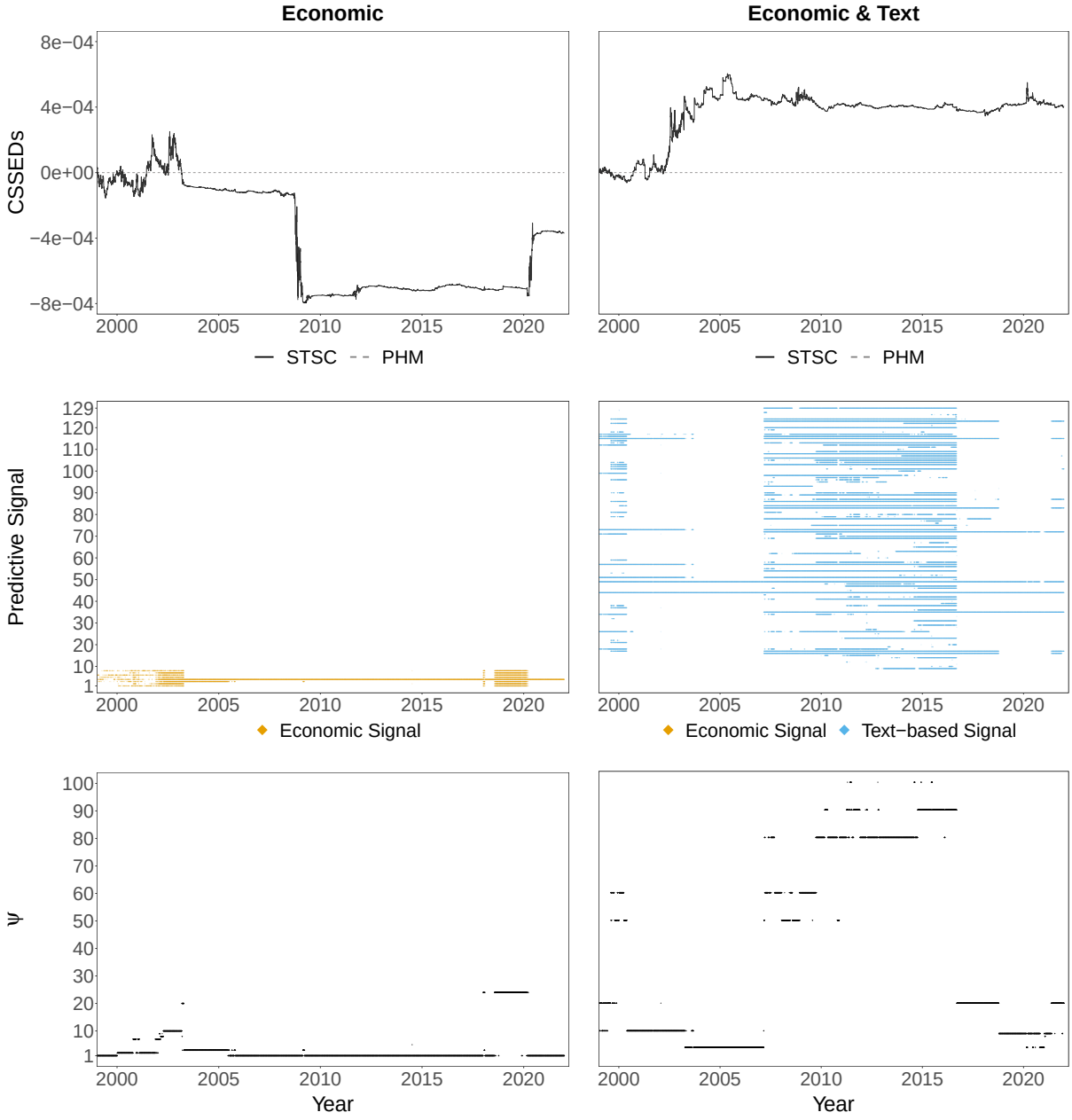

Figure 1: **Forecasting performance, selected predictive signals, and selected subset size for *Economic* and *Economic & Text*.** The top panel shows the cumulative sums of squared error differences between the PHM forecast and STSC. The middle panel depicts which signals were included in STSC over time. The lower panel depicts the evolution of the subset size ψ.

Figure 2 depicts the estimated aggregate predictive densities for the setting *Economics & Text* generated by STSC on the last trading day for each month in 2020. It can be seen that STSC picks up changing conditional volatility, showing an increase at the outbreak of COVID-19 and a subsequent reduction until the end of August. We omit the plots for the setting *Economic* since the aggregate predictive densities look very similar. This is because the observational variance makes up the lion's share of the conditional variance, whereas the uncertainty about the coefficients only accounts for a small part of it.

As a robustness check to evaluate whether STSC can eliminate pure noise predictors, we added 10,000 noise signals generated from the standard normal distribution, leaving our results virtually unchanged. As a further robustness check, we varied one of our tuning parameters in each setup and fixed its value, while leaving the remaining (grids of the) tuning parameters unchanged (see Table 7 in Appendix 6.3). The results are based on the setting *Economic & Text*. The main findings can be summarized as follows: slight deviations from the default value for $\kappa = 0.94$ lead to similar results. A value of κ close to one, however, is detrimental, indicating that (almost) constant volatility is inappropriate, as one would expect. Pure model selection (ψ =1) leads to favorable economic performance, but does not generate a significant CW test statistic. Forcing the subset to high values results in both decreased statistical and economic performance. In particular, the simple average across all available candidate forecasts ($\psi = 36,888$) performs poorly, bolstering the importance of sorting out irrelevant signals. Adopting constant coefficients within the TV-C models ($\lambda = 1$) leads to superior results, both in terms of CW test statistics and economic performance. Overall, the results indicate that the data-driven selection of tuning parameters in STSC works well in the sense that the performance is never much worse than best ex-post choices of tuning parameters. That said, the choices of (grids of) tuning parameters may also be guided by domain-specific knowledge and experiences or recommendations from previous studies (as, for example, our choice of κ). Similarly, we could have restricted λ to 1, arguing that gradual coefficient changes might not be useful for extremely noisy text-based signals. However, to be conservative and to avoid any cherry-picking, we chose the same lower boundary for

λ (in terms of effective window size) across both applications and the simulation study.

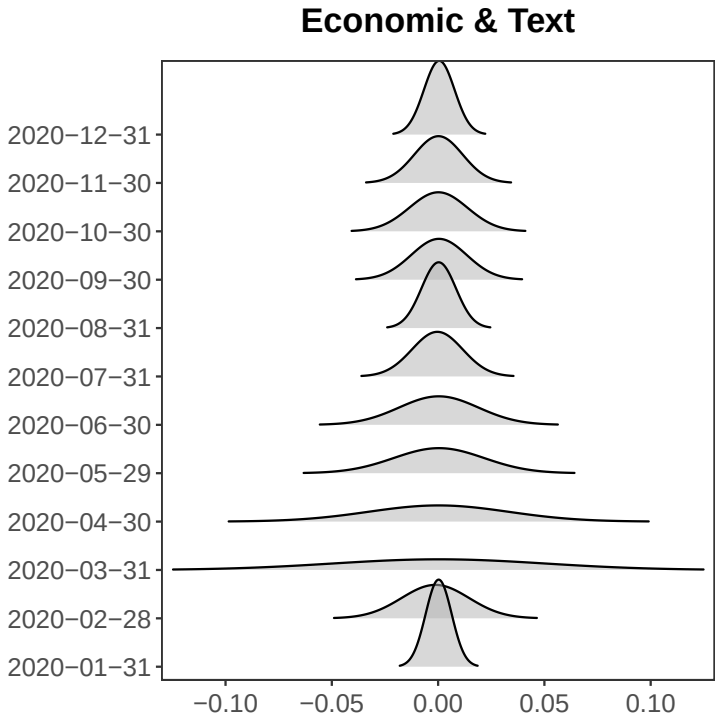

Figure 2: **Estimated aggregate predictive densities for the setting *Economic & Text***. The plot depicts the estimated aggregate predictive densities generated by STSC on the last trading day for each month in 2020.

4.2 Application II: Forecasting quarterly inflation

As one of the key macroeconomic variables, an accurate forecast of price inflation is of paramount interest to policymakers, firms and households alike. Yet, although the forecasting of price inflation has been studied extensively (e.g., Inoue and Kilian, 2008; Faust and Wright, 2013; Koop and Korobilis, 2023), it is still the case that simple approaches such as the autoregressive (AR) model and the unobserved component stochastic volatility model (UC-SV) of Stock and Watson (2007) are hard-to-beat OOS benchmarks in terms of forecasts. We focus on short-term one-quarter-ahead forecasts in our second application.

4.2.1 Data and empirical setup

We consider four measures of US inflation as the target series: GDP deflator (GDPCTPI), PCE deflator (PCECTPI), Total CPI (CPIAUCSL), and Core CPI (CPILFESL). The names in parentheses refer to the variables' codes in the FRED-QD database from McCracken and Ng (2020), from which the data are drawn.

As predictive signals, we use a high-dimensional and heterogeneous data set compiled by Koop and Korobilis (2023), who merge indicators from various macroeconomic and financial sources. In addition to the indicators compiled by McCracken and Ng (2020), the data set comprises portfolio data from Jurado et al. (2015), stock market predictors from Welch and Goyal (2008), survey data from University of Michigan consumer surveys, commodity prices from the World Bank's Pink Sheet database, and key macroeconomic indicators from the Federal Reserve Economic Data for four economies.[18] In total, the data set consists of 441 indicators and can be found here.

We add 16 other predictive signals, which are point forecasts of (the respective measure of) inflation. These point forecasts are based on forecasting models considered in Koop and Korobilis (2023), ranging from simple to highly sophisticated ones. For example, models that use exogenous predictive variables such as Gaussian process regressions, combine multivariate information and can capture nonlinear interactions. The new variational Bayes dynamic variable selection (DVS) approach from Koop and Korobilis (2023) allows for variable selection in high dimensions, and the UC-SV, as well as Bayesian structural breaks AR(2), are specialized techniques for modeling time variation.

We use the same evaluation period as in Koop and Korobilis (2023): the data span the period 1960Q1 to 2021Q4, with OOS evaluations starting in 1991Q2. In sum, with three possible values for λ, we have $441 + 16 = 457$ signals and $(441 + 16) \times 3 = 1,371$ candidate density forecasts.

[18] For further details on the data, see Koop and Korobilis (2023) and the references therein.

4.2.2 Results

Table 4 summarizes the results, where we evaluate STSC and the benchmark methods in terms of the relative mean squared error to the AR(2) model. Values below one indicate more accurate performance in a mean squared error sense and stars indicate statistical significance based on the Diebold and Mariano (1995) test.

Although STSC does not yield the most accurate OOS forecast in each case, it is the only method that is consistently more accurate than the AR(2) model across the four measures of inflation. When comparing STSC's performance with the 16 forecasting models considered in Koop and Korobilis (2023) (see Tables 2 and 3 in their paper), STSC, on average, provides the highest forecast accuracy relative to the AR(2) benchmark. Table 8 in Appendix 6.4 shows that our results are robust to different (grids of) tuning parameters. However, restricting the subset size ψ either to a small value, or taking the simple average over all 1,371 models is (on average) detrimental to performance, again confirming the importance of choosing the subset size flexibly. We find the lowest MSE ratios for PCE Deflator and Total CPI across all methods, whereas the MSE ratios for Core CPI are always above one, except when using STSC. In comparison to Total CPI, Core CPI does not include goods and services from the food and energy sectors. In (unreported) descriptive results we find that Total CPI is less autocorrelated and more volatile than Core CPI, which may explain the higher predictability of Total CPI compared to a simple AR(2) model.

Table 4: Forecast evaluation: Application to inflation.

	GDP Deflator	PCE Deflator	Total CPI	Core CPI
STSC	0.95	0.68**	0.94	0.95
BRT	1.09	0.89	0.84	1.31
RF	1.00	0.90	0.80*	1.15
RLasso	0.92	0.72**	0.86	1.03
XGB	0.93	0.93	0.88	1.19

The table reports the mean squared errors of STSC and the benchmark methods relative to the AR(2) model. Values below one indicate better performance. The evaluation period spans 1991Q2 to 2021Q4. One star indicates significance at the 10% level; two stars significance at the 5% level; and three stars significance at the 1% level from one-sided Diebold and Mariano (1995) test statistics.

The top panels of Figures 3 and 4 show the CSSEDs between the AR(2) forecast and STSC, and between the AR(2) forecast and the benchmark methods BRT, RF, RLasso, and XGB. Positive values indicate better performance relative to the AR(2) benchmark. Forecasting gains, in particular, accrued at the time of the Great Recession and the COVID-19 pandemic, corroborating that predictive indicators are useful during periods of crisis, and that the relative predictive accuracy compared to simple benchmarks increases (see, e.g., Beckmann et al., 2020; Medeiros et al., 2021; Koop and Korobilis, 2023). The middle panels of Figures 3 and 4 show which predictive signals were selected over time. In further (unreported) analyses we computed the ten most often selected signals for each measure of inflation. Across all measures, the expected change in prices over the next year, as measured by the University of Michigan (mnemonic: INFEXP), appears consistently in the top ten, corroborating the importance of survey data for predicting inflation. Interestingly, only one point forecast appears in the top ten, namely Dynamic Model Averaging with five principal components for GDP Deflator. Nevertheless, we find that including both the (441) "simple" signals and the (16) point forecasts of inflation as signals leads to superior results overall than using the sets individually.

The lower panels of Figures 3 and 4 indicate the chosen subset size over time. The number of signals selected, their identities and the length of the episodes in which they were selected, vary considerably over time and across measures of inflation—exploiting the flexibility of the STSC method.

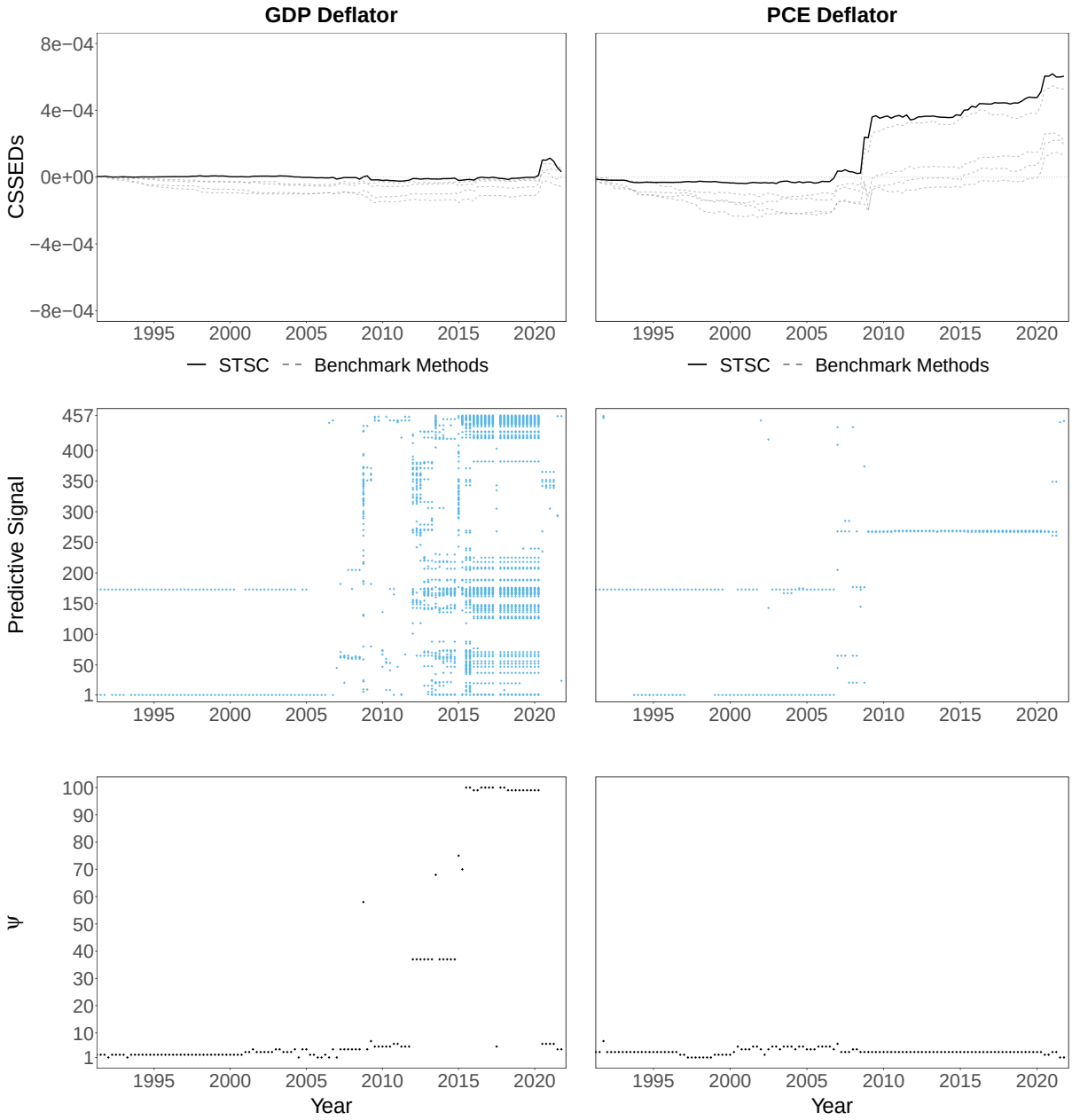

Figure 3: **Forecasting performance, selected signals, and subset size.** The top panel shows the cumulative sums of squared error differences between the AR(2) forecast and STSC, and between the AR(2) forecast and the benchmark methods BRT, RF, RLasso, and XGB. The middle panel shows which signals were included in STSC over time. The lower panel depicts the evolution of the subset size ψ.

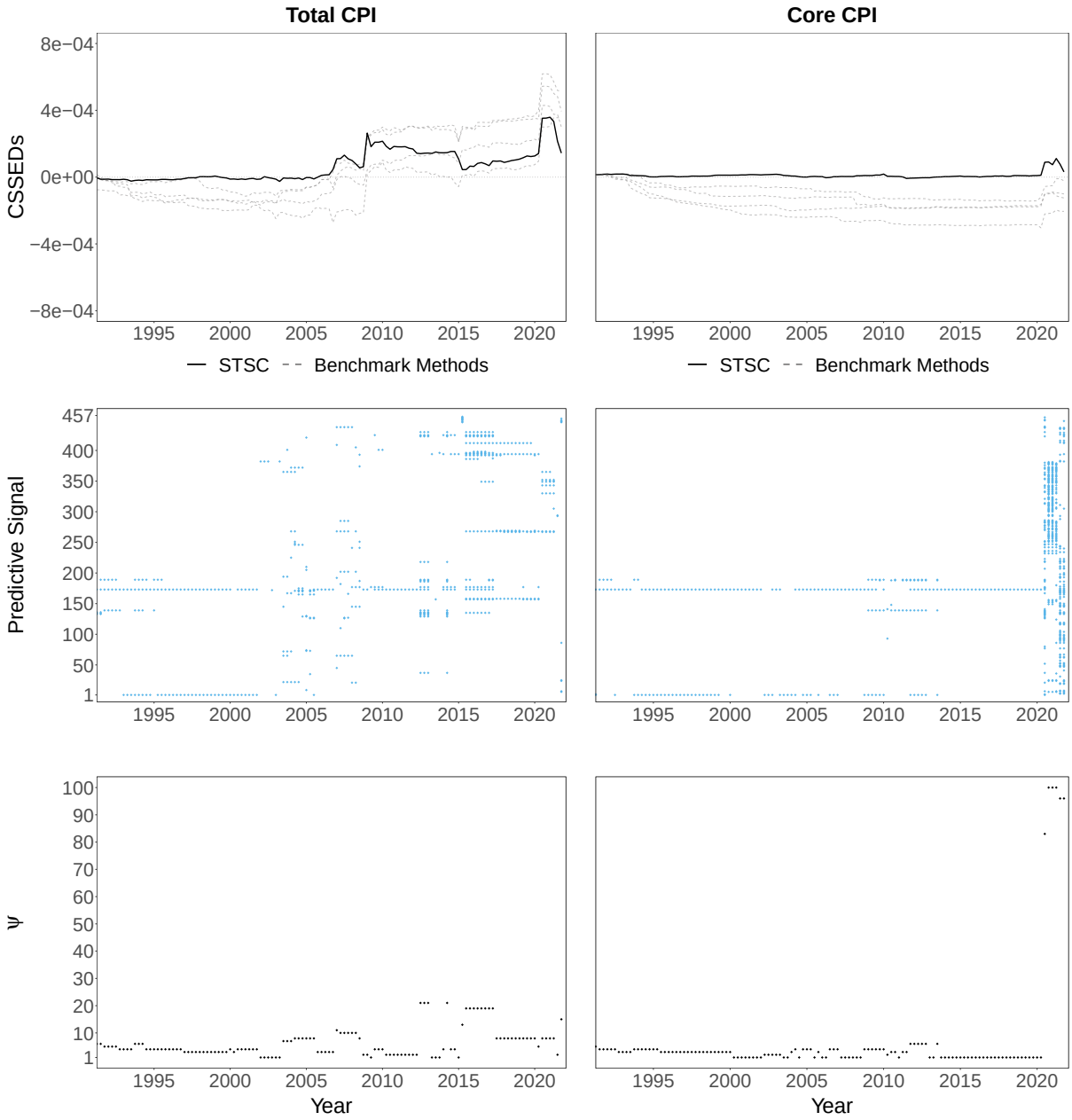

Figure 4: **Forecasting performance, selected signals, and subset size.** The top panel shows the cumulative sums of squared error differences between the AR(2) forecast and STSC, and between the AR(2) forecast and the benchmark methods BRT, RF, RLasso, and XGB. The middle panel shows which signals were included in STSC over time. The lower panel depicts the evolution of the subset size ψ.

The subplots in Figure 5 depict the estimated aggregate predictive densities for the four measures of inflation between 2020Q1 and 2021Q4. The plots illustrate how the estimated conditional mean and volatility in STSC evolve over time. In particular, we can see how the estimated conditional volatility increases in the second half of 2021.

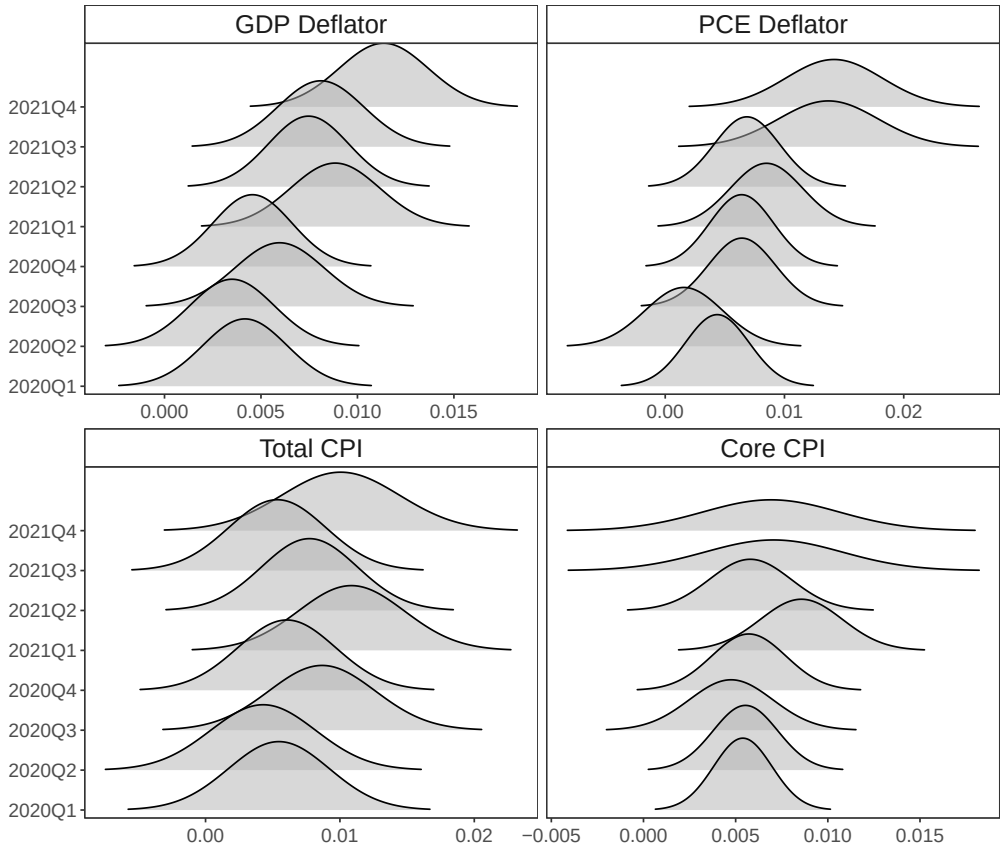

Figure 5: **Estimated aggregate predictive densities.** The plots depict the estimated aggregate predictive densities from 2020Q1 to 2021Q4 for four measures of inflation: GDP Deflator (upper left corner), PCE Deflator (upper right corner), Total CPI (lower left corner), and Core CPI (lower right corner).

5 Concluding Remarks

We have introduced an ensemble learning method for time series forecasting that can handle tens of thousands of predictive signals, many of which are potentially irrelevant or short-lived. In addition to several conceptual advantages of the proposed method, the results of the simulation study and the two applications have shown that our approach provides a versatile tool that has the potential to find its way into the toolbox of applied researchers in time series forecasting. We provide an `R`-package that implements our approach, making it easy for other researchers and practitioners to apply our method to their forecasting problem at hand.

References

Adrian, T., Crump, R. K., and Vogt, E. (2019). Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds. *Journal of Finance*, 74(4):1931–1973.

Aiolfi, M. and Favero, C. A. (2005). Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting*, 24(4):233–254.

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

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, 36(1):197–200.

Barbaglia, L., Consoli, S., and Manzan, S. (2022). Forecasting with economic news. *Journal of Business & Economic Statistics*, pages 1–12.

Beckmann, J., Koop, G., Korobilis, D., and Schüssler, R. A. (2020). Exchange rate predictability and dynamic bayesian learning. *Journal of Applied Econometrics*, 35(4):410–421.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *J. Open Source Software*, 3(30):774.

Bernaciak, D. and Griffin, J. E. (2022). A loss discounting framework for model averaging and selection in time series models. *arXiv preprint arXiv:2201.12045*.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089.

Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232.

Borup, D., Eriksen, J. N., Kjær, M. M., and Thyrsgaard, M. (2023). Predicting bond return predictability. *Management Science*, forthcoming.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Cederburg, S., Johnson, T. L., and O'Doherty, M. S. (2023). On the economic significance of stock return predictability. *Review of Finance*, 27(2):619–657.

Chan, F. and Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting*, 34(1):64–74.

Chen, B. and Maung, K. (2023). Time-varying forecast combination for high-dimensional data. *Journal of Econometrics*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

Clark, T. E. and Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, 30(4):551–575.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

Cogley, T. and Sargent, T. J. (2005). Drifts and volatilities: monetary policies and outcomes in the post wwii us. *Review of Economic dynamics*, 8(2):262–302.

Cong, L. W., Liang, T., and Zhang, X. (2019). Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. *Interpretable, and Data-driven Approach to Analyzing Unstructured Information*, Working paper, Chicago Booth School of Business and Cornell University.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157–181.

Del Negro, M., Hasegawa, R. B., and Schorfheide, F. (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics*, 192(2):391–405.

Demetrescu, M., Georgiev, I., Rodrigues, P. M., and Taylor, A. R. (2022). Testing for episodic predictability in stock returns. *Journal of Econometrics*, 227(1):85–113.

Deutsch, M., Granger, C. W., and Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting*, 10(1):47–57.

Diebold, F. X. (1991). A note on bayesian forecast combination procedures. In *Economic Structural Change: Analysis and Forecasting*, pages 225–232. Springer.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Diebold, F. X. and Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, 35(4):1679–1691.

Diebold, F. X., Shin, M., and Zhang, B. (2022). On the aggregation of probability assessments: Regularized mixtures of predictive densities for eurozone inflation and real interest rates. *Journal of Econometrics*.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40:1–18.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, 71:1054–1063.

Elliott, G. and Timmermann, A. (2005). Optimal forecast combination under regime switching. *International Economic Review*, 46(4):1081–1102.

Faria, A. and Mubwandarikwa, E. (2008). The geometric combination of bayesian forecasting models. *Journal of Forecasting*, 27(6):519–535.

Farmer, L., Schmidt, L., and Timmermann, A. (2023). Pockets of predictability. *Journal of Finance*, forthcoming.

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In *Handbook of Economic Forecasting*, volume 2, pages 2–56. Elsevier.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational s Statistics & Data Analysis*, 38(4):367–378.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620.

Goyal, A., Welch, I., and Zafirov, A. (2023). A comprehensive look at the empirical performance of equity premium prediction ii. *Available at SSRN 3929119*.

Grushka-Cockayne, Y., Jose, V. R. R., and Lichtendahl Jr, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, 63(4):1110–1130.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592.

Hill, R. A. and Rodrigues, P. M. (2022). Forgetting approaches to improve forecasting. *Journal of Forecasting*, 41(7):1356–1371.

Huang, H. and Lee, T.-H. (2010). To combine forecasts or to combine information? *Econometric Reviews*, 29(5-6):534–570.

Inoue, A. and Kilian, L. (2008). How useful is bagging in forecasting economic time series? a case study of us consumer price inflation. *Journal of the American Statistical Association*, 103(482):511–522.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3):1177–1216.

Kang, Y., Cao, W., Petropoulos, F., and Li, F. (2022). Forecast with forecasts: Diversity matters. *European Journal of Operational Research*, 301(1):180–190.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Koop, G. and Korobilis, D. (2023). Bayesian dynamic variable selection in high dimensions. *International Economic Review*.

Lima, L. R. and Godeiro, L. L. (2023). Equity-premium prediction: Attention is all you need. *Journal of Applied Econometrics*, 38(1):105–122.

McAlinn, K. and West, M. (2019). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, 210(1):155–169.

McCracken, M. and Ng, S. (2020). Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., and 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.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Mele, A. (2007). Asymmetric stock market volatility and the cyclical behavior of expected returns. *Journal of Financial Economics*, 86(2):446–478.

Paye, B. S. and Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, 13(3):274–315.

Pettenuzzo, D., Timmermann, A., and Valkanov, R. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics*, 114(3):517–553.

Raftery, A. E., Kárnỳ, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66.

Reuters, J. M. (1996). Riskmetrics-technical document. Technical report, Technical report, JP Morgan-Reuters.

Roccazzella, F., Gambetti, P., and Vrins, F. (2022). Optimal and robust combination of forecasts via constrained optimization and shrinkage. *International Journal of Forecasting*, 38(1):97–116.

Samuels, J. D. and Sekkel, R. M. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, 33(1):48–60.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41(3):788–829.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of forecasting*, 23(6):405–430.

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

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, 1:135–196.

Tsiakas, I., Li, J., and Zhang, H. (2020). Equity premium prediction and the state of the economy. *Journal of Empirical Finance*, 58:75–95.

Wang, X., Kang, Y., Petropoulos, F., and Li, F. (2022). The uncertainty estimation of feature-based forecast combinations. *Journal of the Operational Research Society*, 73(5):979–993.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

West, M. and Harrison, J. (1997). Bayesian forecasting and dynamic models. Springer, 2nd edn.

Yan, X. S. and Zheng, L. (2017). Fundamental analysis and the cross-section of stock returns: A data-mining approach. *The Review of Financial Studies*, 30(4):1382–1423.

Zellner, A., Keuzenkamp, H. A., and McAleer, M. (2002). *Simplicity, inference and modelling: keeping it sophisticatedly simple*. Cambridge University Press.

6 Appendix

6.1 Addition to Application I: Selection of predictive signals

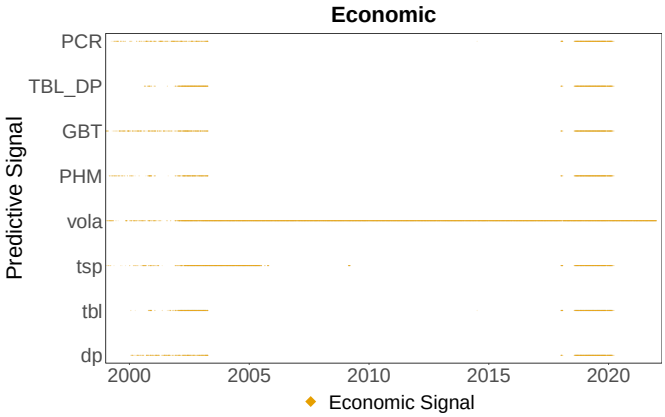

Figure 6: This figure is an enlarged version of the left middle panel in Figure 1 and depicts which signals were included in STSC over time.

Table 5: Selected signals in *Economic & Text*.

crow	hardwood	cushman	bruno	kay	yorktown	econometr	birthday	monetarist	royc
khartoum	privileg	wakefield	cont	crisscross	unreport	emissari	usa	banish	seneg
seclud	ramada	savannah	kravi	excav	waterg	plantat	prerequisit	pornographi	wisdom
panicki	standbi	conduit	intox	hanna	wade	ralph	klerk	pessim	blow
poetri	destabil	seeth	bonus	pragu	platform	loneli	walker	twin	quash
ontario	taper	justifi	oslo	shirley	hostess	loos	soil	luggag	clinch
nascent	ziyang	apartheid	editori	link	imprud	region	bacon	honey	tutu
burglari	repetit	lui	haul	millimet	oyster	trolley	ting	op	reflex
appel	earl	buffalo	wider	tireless	south	monasteri	civil	breather	heart
wale	rohatyn	turbin	craft	mover	fireplac	donor	wood	muscovit	setback
guatemala	warehous	perez	cruiser	tip	loudest	vulner	metropolitan	brick	canal
hector	dakota	calib	tribe	sioux	snarl	midway	alfonso	charit	furi
novak									

The table lists all text-based signals which were selected at least once from *Economic & Text* (in total: 121). The signals are arranged in descending frequency (row-wise).

6.2 Addition to Application I: Low-dimensional setting

We investigate the ability of STSC to detect local predictability in a low-dimensional setting and over an extended evaluation period. To do so, we restrict our set of predictive indicators to four economic signals for which comparatively long data histories are available: *tbl*, *dp*, *tsp*, *vola*. This set of predictive signals was also used by Farmer et al. (2023) to investigate local predictability, but using kernel regression instead. We add the PHM as the fifth signal.

The signals are available from several starting dates. As our approach can easily handle predictive signals of different lengths, we use the earliest possible starting date for each series. After an initial training sample of five years for the candidate forecasts, based on the signal with the shortest data series, our OOS evaluation period spans 1967-07-03 to 2021-12-31.

To investigate STSC's ability to detect local predictability, we compare two setups: in the first one, we run STSC separately for each of the four signals *tbl*, *dp*, *tsp*, *vola*, where STSC can switch between the given signal and the PHM. Similarly, we use STSC with all signals, called *Economic Long* hereafter, where STSC can choose between all signals flexibly. In our second setup, called *no-switch* hereafter, we compute forecasts without a switching option, where each time, the density forecast is based solely on one particular signal.[19] For *Economic Long*, we compute the simple average of all five forecasts for the *no-switch* setup.

Table 6 summarizes the results. For all four signals, STSC, which allows for switching, achieves substantially better evaluation metrics than *no-switch*, indicating that STSC successfully captures local predictability. Similarly, STSC outperforms *no-switch* (i.e., the simple average) in the case of *Economic Long*. As STSC always outperforms *no-switch*, the results in Table 6 indicate that STSC successfully captures local predictive power. We next look at how the performance developed over time.

[19] We set $\lambda = 1$ in *no-switch*.

Table 6: Summary of results for the low-dimensional setting.

	CW	$\widehat{\alpha}$	CER
tbl			
STSC	5.03^{***}	$5.47\%^{***}$	10.32%
no-switch	2.96^{***}	$3.74\%^{***}$	8.92%
dp			
STSC	4.02^{***}	$7.81\%^{***}$	12.90%
no-switch	0.18	-0.42%	5.28%
tsp			
STSC	5.24^{***}	$8.10\%^{***}$	13.12%
no-switch	2.02^{**}	$2.56\%^{**}$	7.91%
vola			
STSC	2.53^{***}	$11.25\%^{***}$	16.62%
no-switch	0.50	0.55%	6.62%

The table reports the Clark and West (2007) (CW) test statistics for OOS predictability measured relative to the PHM. As measures of economic predictability, we report the estimated annualized alpha ($\hat{\alpha}$) and the (annualized) certainty equivalent return (CER) values. One star indicates significance at the 10% level; two stars at the 5% level; and three stars at the 1% level (for one-sided alternatives). The evaluation sample spans 1967-07-03 to 2021-12-31.

For each of the four signals, Figure 7 shows the evolution of cumulative sums of squared error differences (CSSEDs) between the PHM and STSC (black lines), and between the PHM and *no-switch* (dotted gray lines). Positive CSSEDs indicate outperformance of STSC against the PHM.

We observe two striking results. First, the outperformance of the setups that allow for switching are always higher than those which do not, visually demonstrating the benefit of exploiting local predictability. Second, the outperformance against the PHM essentially accrued until the 1980s. After then, the CSSEDs of the *no-switch* strategies suggest that the economic signals were barely useful anymore. The pattern of decreasing predictive power of the economic signals aligns with Farmer et al. (2023) (see Figure IA.1 in their paper) and Demetrescu et al. (2022). Both the CW test statistics and the alphas generated by STSC compare well with those in Farmer et al. (2023) (see Table III in their paper), based on their nonparametric kernel regressions, further strengthening the

credibility of STSC to pick up "pockets" of predictability. Also in line with Farmer et al. (2023), we do not find a higher level of predictability during the NBER recessions in the unreported results.

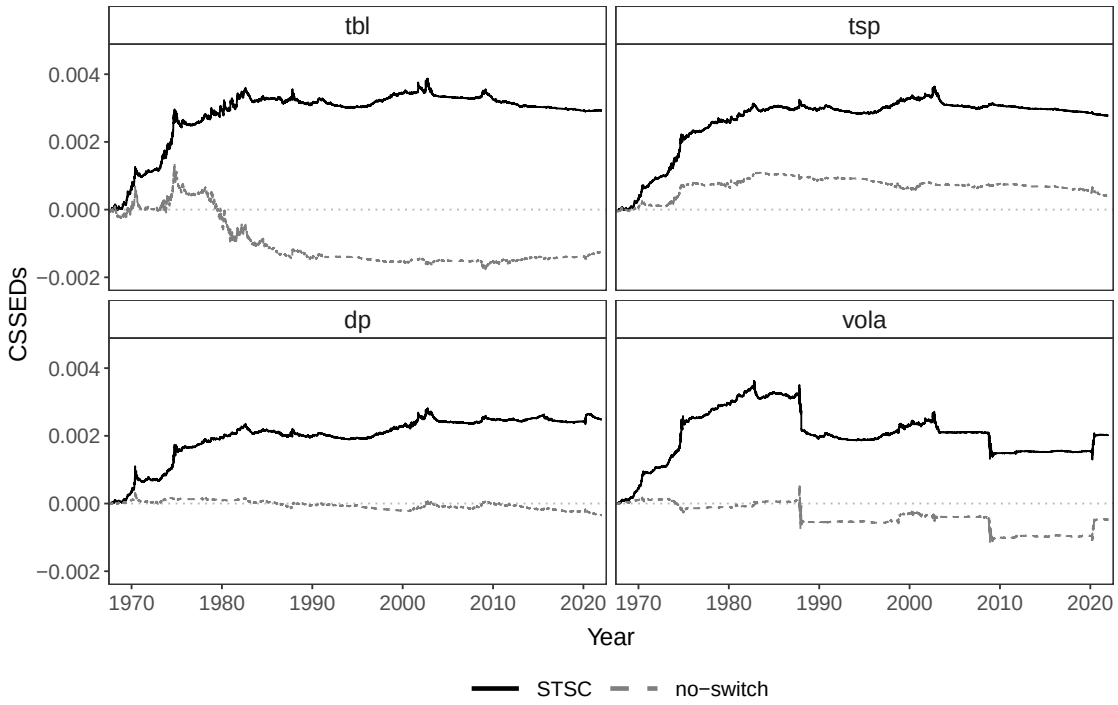

Figure 7: **Cumulative sums of squared error differences between the PHM and the forecasts based on one of the economic signals.** Each plot depicts the cumulative sum of squared forecast error differences between between the PHM and STSC, and between the PHM and *no-switch*.

For *Economic Long*, the top panel of Figure 8 depicts the CSSEDs between the PHM and STSC, and between the PHM and *no-switch*, which equals the simple average of all forecasts. Again, while we observe strong outperformance of STSC against the simple average, beating the PHM in terms of point forecast accuracy became more difficult since the 1980s. The bottom panel of Figure 8 depicts the evolution of the subset size ψ. In the earlier part of the sample, only one forecast was selected each day, while the middle panel of Figure 8 reveals that these forecasts were based on changing signals.

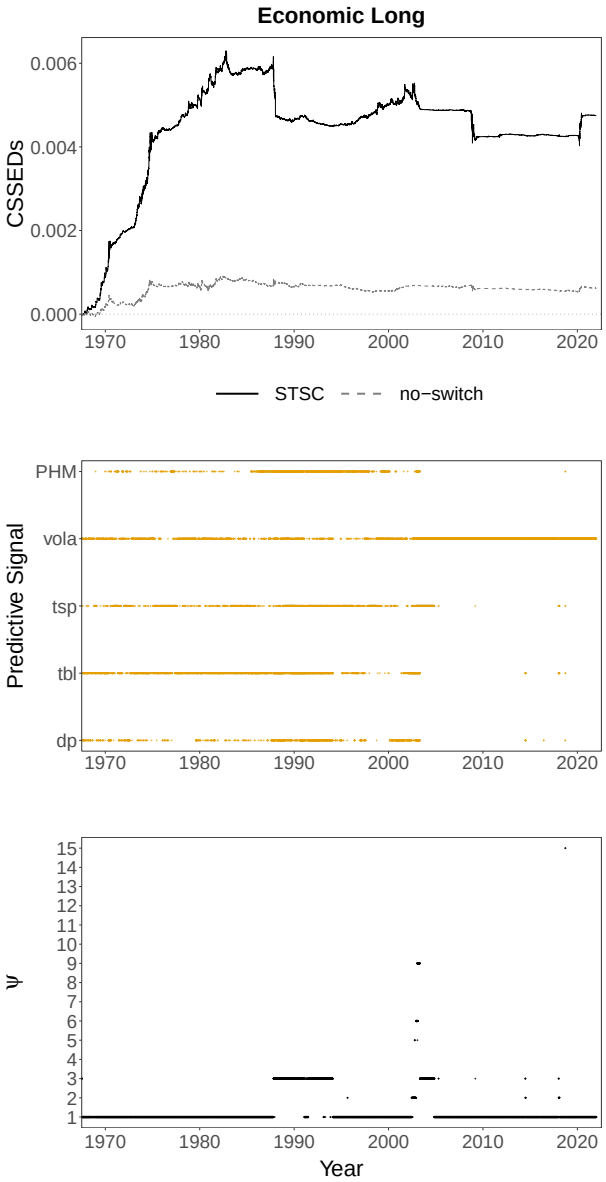

Figure 8: **Summary statistics for *Economic Long*.** The top panel shows the CSSEDs between the PHM and STSC, and between the PHM and *no-switch*, which equals the simple average of all forecasts. The middle panel shows which signals were included in STSC. The lower panel depicts the evolution of the subset size ψ.

6.3 Robustness checks: Application I

Table 7: Robustness results for Application I.

	CW	$\widehat{\alpha}$	CER
Varying κ			
$\kappa = 1.00$	-2.07	1.18%	1.15%
$\kappa = 0.99$	-0.90	1.92%	4.15%
$\kappa = 0.98$	0.42	3.03%	5.17%
$\kappa = 0.97$	0.79	3.16%*	5.24%
$\kappa = 0.96$	1.06	3.77%*	5.75%
$\kappa = 0.95$	1.59*	4.32%**	6.19%
$\kappa = 0.94$	1.48*	4.23%**	6.08%
$\kappa = 0.93$	1.36*	4.58%**	6.42%
$\kappa = 0.92$	1.12	4.25%**	6.09%
$\kappa = 0.91$	1.49*	4.58%**	6.39%
$\kappa = 0.90$	1.23	4.10%**	5.89%
Varying δ			
$\delta = 1.0000$	1.23	4.70%**	6.58%
$\delta = 0.9992$	1.48*	4.23%*	6.08%
$\delta = 0.9984$	1.73**	4.48%**	6.33%
Varying ψ			
$\psi = 1$	0.64	5.68%***	7.35%
$\psi = 5$	1.41*	5.30%***	7.03%
$\psi = 10$	0.89	4.69%**	6.49%
$\psi = 25$	0.94	4.11%**	5.99%
$\psi = 50$	0.79	3.40%*	5.37%
$\psi = 75$	0.63	3.30%*	5.28%
$\psi = 100$	0.50	3.11%*	5.10%
$\psi = 36,888$	-0.30	2.17%	4.22%
Varying λ			
$\lambda = 1.0000$	2.57***	5.61%***	7.77%
$\lambda = 0.9992$	1.50*	3.95%**	5.93%
$\lambda = 0.9984$	1.29*	4.90%**	6.51%

The table summarizes the robustness results for STSC with respect to various choices of (grids of) tuning parameters. In each setup, we varied one of the tuning parameters and fixed its value, while leaving the remaining (grids of) tuning parameters unchanged. One star indicates significance at the 10% level; two stars at the 5% level; and three stars at the 1% level (for one-sided alternatives).

6.4 Robustness checks: Application II

Table 8: Robustness results for Application II.

	GDP Deflator	PCE Deflator	Total CPI	Core CPI
Varying κ				
$\kappa = 1.00$	0.91	0.80**	0.89	0.88
$\kappa = 0.99$	1.04	0.75**	0.89	1.07
$\kappa = 0.98$	0.95	0.68**	0.89	0.95
$\kappa = 0.97$	0.96	0.73**	0.89	0.86
$\kappa = 0.96$	0.85	0.75**	0.89	0.85
$\kappa = 0.95$	0.82	0.78**	0.89	0.87
$\kappa = 0.94$	0.84	0.71***	0.89	0.82
$\kappa = 0.93$	0.86	0.74***	0.89	0.83
$\kappa = 0.92$	0.85	0.75**	0.89	0.81
$\kappa = 0.91$	0.78*	0.78**	0.89	0.82
$\kappa = 0.90$	0.79	0.72***	0.89	0.81
Varying δ				
$\delta = 1.00$	0.93	0.69**	0.99	0.83
$\delta = 0.95$	0.95	0.68**	0.94	0.95
$\delta = 0.90$	0.92	0.71**	0.91	0.98
Varying ψ				
$\psi = 1$	1.10	0.76**	1.00	0.99
$\psi = 5$	1.01	0.73**	0.92	0.88
$\psi = 10$	0.93	0.75**	0.91	0.94
$\psi = 25$	0.92	0.75**	0.87	0.86
$\psi = 50$	0.90	0.76**	0.86	0.89
$\psi = 75$	0.88	0.78**	0.85*	0.87
$\psi = 100$	0.90	0.78**	0.86	0.89
$\psi = 1,371$	1.25	0.89	0.91	0.96
Varying λ				
$\lambda = 1.00$	0.90*	0.87	0.94	1.08
$\lambda = 0.95$	0.94	0.79**	0.93	1.07
$\lambda = 0.90$	1.08	0.69**	1.02	1.01

The table summarizes the robustness results for STSC with respect to various choices of (grids of) tuning parameters. In each setup, we varied one of the tuning parameters and fixed its value, while leaving the remaining (grids of) tuning parameters unchanged. One star indicates significance at the 10% level; two stars significance at the 5% level; and three stars significance at the 1% level from one-sided Diebold and Mariano (1995) test statistics.