A Forest Full of Risk Forecasts for Managing Volatility*

Onno Kleen André P. Santos Anastasija Tetereva

March 16, 2026

Abstract

We propose a hybrid methodology that integrates forest-based machine learning with long-memory models of stock return volatility. Our approach exploits cross-sectional information in a panel of stocks and allows time-varying parameters that are modeled as nonparametric functions of both firm-specific information and changing market conditions. Empirical results on a panel of 1,131 stocks reveal that covariates such as the VIX, idiosyncratic volatility, and momentum have heterogeneous and economically significant effects on the time-varying parameters. In terms of forecasting accuracy, our hybrid approach outperforms a broad range of time-series models and other machine-learning methods over multiple horizons and volatility regimes. Economically, our enhanced risk forecasts yield higher utility for volatility-managed stock investments and deliver minimum-variance portfolios with significantly lower return volatility and higher Sharpe ratios.

Keywords: Accumulated local effects; cross-sectional heterogeneity; HAR model; local linear forest; volatility regimes.

JEL classification: C32; C53; C55; C58; G17.

*Kleen, Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute, e-mail: kleen@ese.eur.nl; Santos, CUNEF Universidad, e-mail: andre.santos@cunef.edu; Tetereva (corresponding author), Erasmus School of Economics, Erasmus University Rotterdam, Burg. Oudlaan 50, Rotterdam, and Tinbergen Institute, e-mail: tetereva@ese.eur.nl. ORCID iD: 0000-0002-1322-245X. We thank Victor DeMiguel, Andrew Patton, Roberto Renò, and Kevin Sheppard for helpful comments. The paper also benefited from helpful discussions during the 2022 VieCo Conference in Copenhagen, the IAAE 2022 Annual Conference at King's College London, the SoFiE 2022 Annual Conference in Cambridge, the 33rd $(EC)^2$ Conference in Paris, the School of Advanced Science on High Dimensional Modeling in São Paulo, and the 32nd Finance Forum in Pamplona, Spain. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-2811 and EINF-6615. Santos acknowledge financial support from Agencia Estatal de Investigación (grant PID2022-138289NB-I00), Comunidad Autónoma de Madrid (grant 2022-T1/SOC-24167), and FAPESP (grant 2023/01728-0).

1 Introduction

This paper employs the idea underlying random forests (Breiman, 2001) to introduce time-varying parameters in models for forecasting stock return volatility. Our proposed heterogeneous autoregressive (HAR) forest model allows stock-specific state and market-related variables, such as the CBOE Volatility Index (VIX), to influence each stock's dynamics differently. In contrast to conventional regression trees, the leaf nodes of our semiparametric HAR-Forest constitute local linear HAR models instead of simple averages across observations. This structure allows us to capture different states, regime switches, and other nonlinearities in the HAR parameters. Empirically, the HAR-Forest outperforms not only the traditional HAR model but also a broad range of competing specifications proposed in the existing literature.

One building block of our approach is the parsimonious HAR model by Corsi (2009). The HAR model posits that the predicted realized variance (RV) can be modeled as a linear function of its lagged value, the average realized variance over the previous week, and the average realized variance over the previous month. Despite numerous extensions (e.g., Bekaert and Hoerova, 2014; Patton and Sheppard, 2015; Bollerslev et al., 2016, 2018), Audrino et al. (2019) document that none uniformly outperforms the original HAR specification. Moreover, Audrino and Knaus (2016) find that the relevance of different lags of RV varies substantially over time, depending on market conditions. Motivated by these findings, we put forward a hybrid specification that allows the HAR parameters to be state-dependent as a function of both aggregate market conditions and firm-level characteristics.

Our HAR-Forest model employs linear models in each node but allows for complex interactions among a large number of covariates, with the dependence structure captured through decision trees. A practical benefit of including state variables within decision trees is that covariates observed at different frequencies can be accommodated without further adjustments. In this sense, our approach shares the motivation of the MIDAS literature (Ghysels et al., 2005)—incorporating mixed-frequency conditioning information—although the mechanism differs: MIDAS parsimoniously parameterizes high-frequency lag polynomials, whereas our trees select splitting points across

frequency-heterogeneous covariates nonparametrically. Finally, decision trees do not require covariates to be measured on a scale that can be transformed to the RV scale, making it more flexible in terms of the variables that can be included.

We contribute to the ML and financial forecasting literature in three ways. First, we introduce a panel version of local linear forests to realized volatility forecasting, building on Athey et al. (2019) and Friedberg et al. (2020). Second, we make the estimation of local linear trees feasible by replacing individual linear models in the leaf nodes with panel HAR models following Bollerslev et al. (2018), which increases the number of observations in each leaf and allows for deeper trees even under rare market conditions. Third, we adapt the subsampling technique of random forests to accommodate the panel structure of our time series, ensuring that the calendar-time sampling preserves the full cross-sectional dependence structure on each sampled date and provides sufficient observations for common covariates such as the VIX at every tree node. This panel structure also distinguishes our work from Patton and Simsek (2023), who explore tree-based local generalized autoregressive score models. Other examples that exploit a panel structure for risk forecasts include Pakel et al. (2011), Barigozzi et al. (2014), Brownlees (2019), Brownlees and Souza (2021), Conrad et al. (2025), and Freire and Kleen (2022).[1]

Our work is closely related to Li and Tang (2025), who combine linear and ML-based models and show that forecast combination adds value for investors. We differ by nonparametrically modeling the *parameters* of an interpretable HAR model rather than the volatility itself, retaining a simple linear model per stock-day observation that can be analyzed both statistically and economically. In addition, we offer an R package that allows researchers and market practitioners to implement our HAR-Forest model, which improves the replicability and transparency of our approach.[2] Finally, our HAR-Forest model also extends the tree-structured HAR model for individual stocks by Corsi et al. (2012) as we consider a wider set of splitting variables while exploiting cross-sectional information, and estimate a random forest-type setup rather than a single tree.

[1]Beyond volatility forecasting, there is also a wide literature explaining the behavior and predicting the dynamics of asset returns using ML; for example, Kelly et al. (2019); Bali et al. (2020); Gu et al. (2020); Bryzgalova et al. (2020); Guijarro-Ordonez et al. (2025). A growing subset of research explores less conventional applications within the domain of fund analysis, see Li and Rossi (2020), DeMiguel et al. (2023) and Kaniel et al. (2023).

[2]An `R` package implementing the HAR-Forest is available at `[URL removed for review]`.

Our empirical analysis investigates the forecasting of realized volatilities for the largest 500 NYSE/NASDAQ/AMEX stocks by market value from the years 2000 to 2021 with 17 different covariates. The largest 500 stocks are chosen at each estimation period, ensuring that the stocks may vary across estimation periods, which mitigates the risk of survivorship bias and results in an effective total of 1,131 unique stocks considered across all estimation windows. In comparison to Christensen et al. (2023) and Li and Tang (2025), our sample includes the market turbulence of the Covid-19 pandemic in 2020–2021. Moreover, our results obtained with real market data are also supported by a Monte Carlo simulation analysis that studies the finite-sample performance of the HAR-Forest model under cross-sectional heterogeneity and macro-financial regimes observed in real data.

We compare our model with the panel HAR model by Bollerslev et al. (2018) (which is nested in our HAR-Forest specification), the single-tree HAR model by Corsi et al. (2012), individually estimated HAR models with and without realized semi-variances and realized kurtosis (Patton and Sheppard, 2015; Bollerslev et al., 2016), the exponential realized volatility (HExp) model of Bollerslev et al. (2018), conventional random forest models, and a equally-weighted ensemble of all model forecasts. The out-of-sample (OOS) analysis demonstrates that the HAR-Forest consistently outperforms all benchmark models across different forecast horizons. As statistical performance measures, we use loss functions that are robust to measurement error in volatility estimates, such as the squared error (SE) and the QLIKE loss, and calculate utility benefits for risk-targeting investors to put the results into perspective. The MCS procedure by Hansen et al. (2011) confirms that the HAR-Forest performs very well for forecast horizons from 1 day to 66 days ahead.

Overall, our stock-specific results complement the works of Christensen et al. (2023), who analyze the forecast performance of multiple standard machine learning techniques in a smaller cross-section of 29 stocks. Unlike Li and Tang (2025) and our analysis, Christensen et al. (2023) omit overnight returns – even though these returns constitute a significant share of total stock return volatility (see, e.g., French and Roll, 1986) and play a critical role in popular volatility timing strategies (see, e.g., Harvey et al., 2018). Moreover, for a joint forecast evaluation across stocks, we assess the economic performance of our different volatility forecasts in a minimum-variance

portfolio application. To isolate the contribution of volatility forecasting accuracy – abstracting from differences in correlation estimation – we construct covariance matrices by combining each model's variance forecasts with a common correlation matrix that varies over time but is identical across all competing methods. Our results show that the better variance forecasts from our HAR-Forest translate into considerably lower portfolio return volatility and higher Sharpe ratios when compared to standard volatility models.

Studying the accumulated local effects (ALEs) of different covariates reveals the nonparametric dependence of HAR coefficients on the splitting covariates and provides insight into the inner workings of the HAR-Forest model. Our findings reveal substantial differences when comparing the ALE plots across covariates and HAR-Forest coefficients. We detect pronounced changes in the HAR-Forest coefficients for VIX values above 20, suggesting that periods of financial uncertainty are associated with a shorter memory volatility process. This result is consistent with Gallo and Otranto (2015) and Cipollini et al. (2021), who find that turmoil regimes in realized volatility exhibit faster mean reversion than tranquil periods, implying that volatility memory is effectively shorter during episodes of financial stress. Moreover, our variable-importance measure based on the ALEs reveals that lagged aggregated measures of RV, VIX, idiosyncratic volatility, and momentum are among the most important drivers of the coefficients. Finally, when examining the conditional performance of our model across volatility regimes, the HAR-Forest model emerges as superior to individual and panel HAR models. Conventional random forest models, while competitive in lower volatility deciles, struggle to perform well in higher volatility states.

Our findings are supported by an extensive set of robustness checks. The improvements in forecast performance hold across firm characteristics, alternative forecast horizons (up to 66 days), different estimation window lengths, and a range of hyperparameter configurations. Moreover, the HAR-Forest consistently outperforms alternative ML-inspired benchmarks, including regularized models with cross-sectional volatility spillovers and nonlinear specifications such as neural networks.

2 Modeling and forecasting realized variances

Let $r_{t-j\cdot\Delta}$ denote 5-minute log-returns of an asset where $\Delta = 1/M$ and M denotes the number of intraday observations at day t.[3] We forecast the quadratic variation of stock returns in line with the volatility forecasting literature (Andersen and Bollerslev, 1998). As the quadratic variation is not directly observable, we employ the RV defined as the sum of M squared 5-minute log returns as a proxy for the quadratic variation,

$$RV_t^d = \sum_{j=1}^{M} r_{t-(j-1)\cdot\Delta}^2, \tag{1}$$

To get a measure for the stock price variance of the entire day, we include the squared overnight return in RV_t^d.

At time t, we aim to forecast the average RV over the upcoming h days; that is $RV_{t+1:t+h} = \frac{1}{h}\sum_{i=1}^{h} RV_{t+i}^d$.[4] In general, the time series of the RV^d are highly persistent. Figure 1 depicts the RV's empirical autocorrelation functions (ACF) for the subset of stocks in our sample for which we have at least 2500 observations per stock.[5] The autocorrelations are estimated using the instrumental variable estimator suggested by Hansen and Lunde (2014). We employ their preferred specification, a two-stage least-squares estimator with 4 to 10 lagged values of RV as instrumental variables (Hansen and Lunde, 2014). The ACFs of Apple Inc. and Boeing Inc. are depicted in green (dotted line) and blue (dashed line). All other ACFs are in light gray, and the average ACF is in red (solid line). We observe that the persistence in daily RVs varies across stocks. For example, the Apple ACF is particularly higher than the Boeing ACF. One possible explanation is that differences in business models – for instance, firms relying on long-term contracts versus those exposed to high-frequency consumer demand – may contribute to such variation in persistence. This motivates us to relax the pooled panel HAR restrictions in Bollerslev et al. (2018) in our HAR-Forest specification.

[3] For example, on a regular trading day at the New York Stock Exchange (NYSE) we have 78 5-minute returns.

[4] It is known that RV^d is a consistent estimator of the quadratic variation but that it is not robust to microstructure noise (Andersen et al., 2011). However, it has been shown to be a fairly robust choice as a trade-off between using high-frequency data and obstructing micro-structure noise-related estimation errors (Liu et al., 2015; Aït-Sahalia and Xiu, 2019).

[5] The underlying data is described in detail in Section 3.1.

2.1 Heterogeneous autoregressive models for realized variances

Corsi (2009) introduced the HAR model, an additive model with heterogeneous components related to the realized volatility literature. The author motivates the model using the heterogeneous market hypothesis formulated by Müller et al. (1993). Market participants can be classified into three heterogeneous groups of agents according to their trading horizons, i.e., daily, weekly, and monthly. Each group creates its own component of latent volatility. This cascade of time scales in the volatilities motivates the HAR equation:

$$RV_{t+1}^{d} = \overline{RV} + \beta^{d}(RV_{t}^{d} - \overline{RV}) + \beta^{w}(RV_{t}^{w} - \overline{RV}) + \beta^{m}(RV_{t}^{m} - \overline{RV}) + \varepsilon_{t}, \tag{2}$$

where $\beta = (\beta^{d}, \beta^{w}, \beta^{m})^{T}$ is a parameter vector, the weekly and monthly averages of volatility are defined as $RV_{t}^{w} = RV_{t-5:t-2}$ and $RV_{t}^{m} = RV_{t-22:t-6}$. $\overline{RV}$ is the average RV of the stock, and ε_{t} is an innovation term. We write the HAR equation in deviations from the stock-specific long-run variance $\overline{\mathrm{RV}}$ so that pooled estimation across stocks does not require explicit fixed-effect intercepts (see Section 2.3). Finally, predictions for forecast horizons larger than one day can be obtained by substituting RV_{t+1}^{d} with $RV_{t+1:t+h}$ on the left-hand side of Eq. (2); that is, we employ direct forecasts.

Because unconditional distributions of RV are highly right-skewed Andersen et al. (2001), the HAR model is often estimated on logarithmic RV, which yields approximately Gaussian residuals and can improve both in-sample fit and predictive performance (Corsi et al., 2008). Accordingly, we also consider log-RV specifications of the HAR model and of the HAR-Forest; the aggregation of forecasts across trees in the log-RV case is described in Section 2.3.

2.2 Tree-structured HAR models

We investigate how best to incorporate time-varying weights into the HAR model. For example, one approach to incorporating time dependence is to fit models on shorter time series. Similarly, model averaging can be a solution in the presence of structural breaks (Pesaran and Pick, 2011). However, it is difficult to capture both short-term market conditions and regime switches by only

adjusting the size of the estimation window.

An alternative to adjusting the estimation window is to model the parameter variation directly as a function of observable state variables. One natural approach is to consider dynamic HAR models that allow the coefficients to depend on market conditions. However, these models require parametric assumptions about the form of this dependence. One such model is the HARQ model (Bollerslev et al., 2016) that employs the realized quarticity as an interaction term for quantifying the measurement error in daily RV estimates. In addition to the HARQ model, we employ the semi-variance HAR model (SHAR) (Patton and Sheppard, 2015) as another benchmark that employs realized semi-variances instead of total variances. In this model, positive and negative returns can affect future volatility to different degrees. We refer to the Internet Appendix IA.1 for a description of those existing models. Adding a larger number of covariates in such parametric models might be practically feasible using regularization techniques. Nonetheless, regularized regressions still need the data to be transformed to be (approximately) linearly related to RV.

ML approaches, in particular random forests, are natural alternatives to incorporate time-varying coefficients into the HAR model. For that purpose, we shift the focus from predicting the *values* of the realized volatilities by some nonparametric function to modeling *parameters* $\beta = (\beta^d, \beta^w, \beta^m)^T$ in Eq. (2) as a function of, for example, market conditions and firm-specific characteristics. Our forest approach makes it easy to flexibly model evolving parameters in a data-driven way and capture interactions of different regimes with respect to different market and asset-specific conditions.[6] This is in contrast to latent regime-switching models in which only a small number of regimes can be estimated and the number of regimes is difficult to determine.

More precisely, we assume that all splitting rules in tree $\mathcal{T}$ are based on the state of J splitting variables $\mathcal{J}$. The state vector of these J splitting variables is denoted by $Z_t \in \mathbb{R}^J$. Now, $\mathcal{T}$ assigns each possible value of Z_t one of K terminal nodes denoted by $R_1, \ldots, R_K$. The local model in terminal node R_l is given by

$$RV_{t+1:t+h} = \overline{RV} + \beta_l^d(RV_t^d - \overline{RV}) + \beta_l^w(RV_t^w - \overline{RV}) + \beta_l^m(RV_t^m - \overline{RV}) + \varepsilon_t. \tag{3}$$

[6]Bettencourt et al. (2024) employ a similar approach to ours for modeling time-varying portfolio weights.

An example of a tree with local HAR models is given in Figure 2. Here, the set of splitting covariates $\mathcal{J}$ contains weekly and monthly volatilities, i.e., RV_t^w and RV_t^m. The final partitioning contains five regions or a set of five different linear models. Given that the splitting rules of the tree $\mathcal{T}$ are fixed and the final node assignment is determined by the state vector Z_t, the tree implies a mapping $\widetilde{\mathcal{T}}$ so that $\beta_t = \widetilde{\mathcal{T}}(Z_t)$.

The set of splitting covariates can include the independent variables from the HAR model (i.e., RV, RV^w, RV^m) as in Figure 2. It can also include variables characterizing market conditions that may not be included in the linear model itself, e.g., the VIX. Moreover, splitting covariates can also be observed at a lower frequency than daily; for example, monthly stock market betas.

The model in Eq. (3) is estimated in analog to simple classification and regression trees. In every node, the objective is to find the splitting point that minimizes some objective function. Such a minimum is searched over all possible values in all splitting covariates. In other words, the following sum of two residual sum of squares needs to be minimized:

$$\min_{c\in\mathbb{R},\ j\in\mathcal{J}} \left(\min_{\beta} \sum_{t|Z_{j,t}<c} \widehat{\varepsilon}_t^2 + \min_{\beta} \sum_{t|Z_{j,t}\geq c} \widehat{\varepsilon}_t^2 \right), \tag{4}$$

where c is a candidate threshold value drawn from the splitting covariate $Z_{j,t}$, $j = 1, \ldots, J$, and each inner minimization re-estimates the HAR coefficients β separately in the left ($Z_{j,t} < c$) and right ($Z_{j,t} \geq c$) child nodes. Consequently, $\hat{\varepsilon}_t$ in each child denotes the residual from the locally re-fitted HAR model. The computational cost of this optimization is discussed in Section 3.2.

Although a tree of HAR models is more flexible than a single HAR model, it inherits two well-known drawbacks from conventional classification and regression trees. First, the tree needs to be pruned, or an appropriate stopping rule needs to be chosen to avoid data overfitting, which typically requires tuning additional hyperparameters. Second, individual trees are typically unstable because small changes in the data can change the estimated tree significantly.

The HAR-Tree approach above and the HAR-Forest approach below are inherently different from applying conventional regression trees to volatility modeling. For example, Christensen et al. (2023) apply regression trees (among other ML algorithms) to model the realized volatility

as a nonparametric, piecewise constant, function of its past values. Our approach is different. Instead of predicting the realized volatility as a simple average of all observations in the leaf, we predict the realized volatility by a *leaf-specific* HAR model. Through this approach, we employ regime-dependent HAR models whose parameters are subject to regimes governed by changing market conditions and stock-specific characteristics. A concept similar to ours, that is using trees to capture market regimes, was explored in Cong et al. (2024).

2.3 HAR forests

To address the instability of individual trees, we aggregate across trees following the random forest approach. A drawback, however, is that a linear HAR model must be estimated at every leaf. Unlike conventional regression trees, which typically require a minimum node size of only five observations, fitting local linear models accurately demands substantially larger leaf sizes. To circumvent this limitation, we switch from leaf-specific *individual* HAR models to leaf-specific *panel* HAR models as in Bollerslev et al. (2018). Rather than pooling the entire panel for estimation, we use decision trees to capture grouped heterogeneity in volatility behavior.[7] Although the number of days on which the VIX exceeds 50 is fixed, the panel structure increases the number of cross-sectional observations available on such days. It is well known that pooled estimators introduce some bias due to heterogeneity but yield substantial efficiency gains from pooling (e.g. Pesaran and Smith, 1995). We therefore demean RVs in Eqs.(2) and (3), allowing the use of simple pooled estimators without explicit fixed-effect estimation.

The model of each individual tree targeting the average h-step-ahead volatility is

$$RV_{i,t+1:t+h} - \overline{RV_i} = \left(\beta_l^d (RV_{i,t}^d - \overline{RV_i}) + \beta_l^w (RV_{i,t}^w - \overline{RV_i}) + \beta_l^m (RV_{i,t}^m - \overline{RV_i})\right) I_{\mathcal{T}(Z_{\cdot,i,t}) \in R_l} + \varepsilon_{i,t}, \quad (5)$$

where $i = 1, \ldots, N$ with N being the number of assets in the model. Entries in the splitting covariates matrix $Z_{\cdot,\cdot,t} \in \mathbb{R}^{J \times N}$ of all characteristics and stocks at time t may depend on t or/and i like $RV_{i,t}^d$, VIX_t, or $\overline{RV_i}$. In the case of a forest full of panel HAR models, the splitting objective

[7] Using trees to account for heterogeneous return predictability is also employed in Cong et al. (2023) and Cong et al. (2025).

also needs to be slightly adjusted, i.e., the sum of losses of all (pooled) HAR models in the panel has to be considered at every split:

$$\min_{c\in\mathbb{R},\ j\in\mathcal{J}}\left(\min_{\beta}\sum_{i,t|Z_{j,i,t}<c}\widehat{\varepsilon}_{i,t}^2+\min_{\beta}\sum_{i,t|Z_{j,i,t}\geq c}\widehat{\varepsilon}_{i,t}^2\right), \tag{6}$$

where $\widehat{\varepsilon}_{i,t}^2$ are derived from Eq. (5). In the spirit of classical random forests, we fit $b=1,\ldots,B$ individual panel HAR-Trees (Eq. (5)). Then, the coefficients in the HAR-Forest and the HAR-Forest's forecasts are averages of these B coefficients and forecasts. For each HAR-Forest coefficient and stock i, the forest implies mappings $\beta^{\cdot}(z_1,\ldots,z_J):\mathbb{R}^J\to\mathbb{R}$, with $\beta_{i,t}^{\cdot}=\beta^{\cdot}(Z_{\cdot,i,t})$, where $\beta^{\cdot}(Z_{\cdot,i,t})$ is a placeholder for $\beta^d(Z_{\cdot,i,t})$, $\beta^w(Z_{\cdot,i,t})$, or $\beta^m(Z_{\cdot,i,t})$. In the following, we omit $Z_{\cdot,i,t}$ as a function argument if this facilitates readability.

Finally, when the HAR-Forest is estimated on log-RV, aggregation across trees uses the median forecast rather than the mean. To see why, let $\hat{f}_{b,i,t}$ denote the log-RV forecast of tree b for stock i at time t. If the log-forecasts are approximately normally distributed across trees with mean $\mu_{i,t}$ and variance $\sigma_{i,t}^2$, the level-forecasts $\exp\big(\hat{f}_{b,i,t}\big)$ are approximately log-normally distributed, and mean aggregation yields an expected value of $\exp\big(\mu_{i,t}+\sigma_{i,t}^2/2\big)$. This exceeds the median, $\exp(\mu_{i,t})$, by a multiplicative factor of $\exp\big(\sigma_{i,t}^2/2\big)$ that grows with cross-tree disagreement. The median, being equivariant under the monotone transformation $\exp(\cdot)$, satisfies $\text{median}_b\{\exp\big(\hat{f}_{b,i,t}\big)\}=\exp\big(\text{median}_b\{\hat{f}_{b,i,t}\}\big)$ and is free of this bias.

2.3.1 Estimation of HAR forests

Subagging: The estimation scheme of the HAR-Forest follows the general idea of honest random forests (Wager and Athey, 2018), but we adjust the subagging (subsample aggregating) procedure to accommodate our panel structure. Given the panel structure of our data, we employ calendar-time sampling rather than sampling randomly from the entire panel. Our reasoning for this sampling is easily illustrated when including the VIX as a covariate. Note that the VIX has the same value for all stocks on a given day. The optimal split is determined via a grid search over all possible VIX values in each nodes' subsample. With this in mind, our calendar-time sampling

procedure guarantees that we have enough observations per VIX value at each node to fit a local linear model, and, as a byproduct, we preserve the full cross-sectional structure at each node. To save computational time, when the number of unique values of a splitting variable exceeds 91 (corresponding to the number of grid points in our percentile grid $\{0.05, 0.06, \ldots, 0.95\}$), we restrict the search to the 5th through 95th percentiles.

Feature subsampling: Breiman (2001) proposes to decorrelate the trees that form a forest by looking at subsamples of features for each split. We follow the arguments of Breiman (2001) and Goulet Coulombe (2024), and draw random subsets of features $\mathcal{J}^- \subset \mathcal{J}$ with subsampling rate 1/3 instead of $\mathcal{J}$ in Eq. (6). In contrast to Friedberg et al. (2020) and Goulet Coulombe (2024), we refrain from including a ridge regression penalty term in Eq. (6). The reason is that we have relatively large sample sizes in our final leaves with minimum sizes of 100 observations which we consider sufficient to obtain good OLS estimates for our local HAR models. Across estimation periods, aggregating 200 trees per forest is more than sufficient to produce stable out-of-sample forecasting results. The robustness of our results to the choice of hyperparameters is discussed in Sections IA.7.1–IA.7.3 of the Internet Appendix.[8]

2.3.2 Interpretability of HAR forests

To gauge the behavior of the HAR-Forest coefficients over time and their dependence on covariates, we apply a model-agnostic technique called accumulated local effects (ALE) plots introduced by Apley and Zhu (2020). Intuitively, ALE plots measure how a small change in one splitting covariate shifts a HAR-Forest coefficient, while averaging out the influence of all other covariates. Let $z_{(x;-j),i,t}$ be the altered characteristics vector equal to $z_{\cdot,i,t}$, in which the jth element is replaced by x. In contrast to the permutation-based variable importance measures of Fisher et al. (2019), which may be sensitive to multicollinearity among predictors, the ALE-based approach adopted here is more robust in the presence of highly correlated covariates The ALE function for stock i with respect to

[8]In line with our results, random forest-type methods are often found to be relatively insensitive to hyperparameter choices. In contrast, studies such as Mücher (2021) highlight the importance of careful specification of architecture and parameter tuning for neural networks.

splitting variable j and local coefficient $\beta^{\cdot}$ is defined as:

$$ALE_{j,i}^{\beta^{\cdot}}(z) = c + \int_{\min(z_{j,i,t})}^{z} \mathbf{E}\Big[\frac{\partial \beta^{\cdot}}{\partial z_j}\big(Z_{(x;-j),i,t}\big) \;\Big|\; Z_{j,i,t} = x\Big]\,\mathrm{d}x.$$

The constant c is chosen so that the average local effect equals zero under the marginal distribution of $Z_{j,i,t}$. The key distinction from partial dependence plots is that the partial derivative is averaged conditionally on $Z_{j,i,t}$=x, not marginally, which avoids distortions from correlated covariates. We assume the state-dependence of our coefficient vector to be sufficiently smooth so that the ALE functions are well defined. The role of the partial derivative $\frac{\partial \beta^{\cdot}}{\partial z_j}$ is to capture the local effect of variable j. Empirically, ALE functions before standardization can be approximated by splitting the values of a conditional covariate into equally-spaced bins $(\tilde{z}_0, \tilde{z}_1], \ldots, (\tilde{z}_{L-1}, \tilde{z}_L]$:

$$\widetilde{\widehat{ALE}}_{j,i}^{\beta^{\cdot}}(z) = \sum_{l:\tilde{z}_l \le z} \frac{1}{n_l} \sum_{t:z_{j,i,t} \in (\tilde{z}_{l-1}, \tilde{z}_l]} \Big[\beta^{\cdot}\big(z_{(\tilde{z}_l;-j),i,t}\big) - \beta^{\cdot}\big(z_{(\tilde{z}_{l-1};-j),i,t}\big)\Big],$$

where n_l is the number of observations for which $Z_{j,i,t}$ is in the interval $(\tilde{z}_{l-1}, \tilde{z}_l]$. The centered ALE function is then simply given by $\widehat{ALE}_{j,i}^{\beta^{\cdot}}(z) = \widetilde{\widehat{ALE}}_{j,i}^{\beta^{\cdot}}(z) - \frac{1}{L}\sum_{l=1}^{L} \widetilde{\widehat{ALE}}_{j,i}^{\beta^{\cdot}}(z_l)$.

In our context, ALE plots are more suitable than partial dependence plots because of the high correlation among covariates in our financial application. It is important to note that conventional ALE plots illustrate the change in predictions for different values of the feature of interest. In our application, we compute the change in the coefficients of the state-implied HAR model rather than the change in the predicted RV.

2.3.3 Simulation study

To assess finite-sample performance, we conduct a Monte Carlo simulation study comparing the HAR-Forest with the individual HAR and panel-HAR benchmarks. Section IA.3 of the Internet Appendix presents the detailed simulation design, including the data-generating process, estimation setup, and comparative performance metrics. The data-generating process features a locally HAR but globally nonlinear panel structure with cross-sectional heterogeneity across two groups of

stocks and macro-financial regime dependence driven by the historical VIX path. The daily HAR coefficient depends nonlinearly on lagged realized volatility, the weekly coefficient is modulated by the VIX, and both vary across groups, thereby reproducing key stylized facts observed in financial data. The individual HAR and panel HAR deliver near-identical forecast accuracy, as expected given the time-varying (rather than stock-fixed) nature of the heterogeneity. In contrast, the HAR-Forest achieves systematically lower MSE across all panel sizes ($N \in \{10, 20, 40\}$), with two-sample t-tests confirming statistically significant gains ($p < 0.05$) in every configuration. The advantage grows with the size of the cross-section: for the largest panel ($N = 40$), the HAR-Forest achieves MSE ratios of 0.95–0.97 relative to the individual HAR, highlighting the importance of locally adaptive persistence dynamics when heterogeneity and regime dependence are pronounced.

3 Empirical analysis

3.1 Data

Our data set contains a range of variables for the 500 largest NYSE/NASDAQ/AMEX stocks by market value, between the years 2000 to 2021 that are classified as common stocks in the Center for Research in Security Prices (CRSP). The largest 500 stocks are chosen at each estimation period, ensuring that the stocks may vary across periods, which mitigates the risk of survivorship bias and results in an effective total of 1,131 unique stocks considered across all estimation windows. Similar to Bollerslev et al. (2019) and Bollerslev et al. (2022), we merge daily CRSP data with NYSE Trade and Quote (TAQ) intraday data. We obtain daily open and close prices from the daily CRSP data files and intraday trade data from the NYSE TAQ database. We merge the two data sets via the Wharton Research Data Services (WRDS) linking tables. The intraday data is cleaned according to Barndorff-Nielsen et al. (2009), and we include only trades from the exchange that is referenced in the daily CRSP data. In line with our definitions for the decorrelated RV cascade (Eq. (1)), we define the daily, weekly, and monthly returns as $Ret^d_{i,t} = \sum_{j=1}^{M} r_{i,t-(j-1)\cdot\Delta}$, $Ret^w_{i,t} = Ret_{i,t-5:t-2}$, $Ret^m_{i,t} = Ret_{i,t-22:t-6}$, where M is the number of intraday observations at day t.

We include several potential splitting variables in our local linear forest. First, we include

lagged returns as splitting variables to capture the leverage effect in our forests. Second, we include the RV^d, RV^w, and RV^m that constitute the right-hand-side equation in the individual HAR models. Third, we include additional daily stock-specific information derived from intraday data: semi-variances of positive and negative returns denoted by RV^+ and RV^- (Barndorff-Nielsen et al., 2010; Patton and Sheppard, 2015) and a jump-robust measure of volatility denoted by MedRV (Andersen et al., 2012). Higher-order moments included in this analysis are the realized skewness and realized kurtosis (Amaya et al., 2015). Section IA.2 of the Internet Appendix provides detailed definitions of these additional covariates.

The VIX is also a splitting variable common to all stocks at each day t. It has been employed in HAR models since Bekaert and Hoerova (2014), who forecast aggregate stock market volatility instead of individual stocks. Monthly idiosyncratic volatility, beta, 12-month momentum, size, and dollar volume traded are taken from Gu et al. (2020) and merged via the CRSP PERMNO identifier.

Table 1 reports cross-sectional summary statistics for all splitting variables. For each variable, we compute stock-specific statistics over the full sample period 2000–2021 and report their cross-sectional averages across the 1,131 stocks in our dataset. The statistics reveal substantial heterogeneity across stocks in the level, dispersion, and higher-order moments of realized volatility and its components, motivating the use of stock-specific time-varying parameters in our HAR-Forest specification.

3.2 Implementation details

Estimation scheme: We employ a rolling window estimation scheme with yearly re-estimation. The first estimation sample comprises five years of daily data (2000–2004), and all models are re-estimated annually on a rolling five-year window, yielding out-of-sample forecasts from 2005 through 2021. We employ the rolling window scheme for two reasons: First, it accounts for possible fundamental changes (i.e. if the average RV of a stock changes over time). Second, we can employ Diebold-Mariano-type tests for forecast comparison even though the models are nested (Diebold and Mariano, 1995; Giacomini and White, 2006). However, Chassot and Audrino (2025) find that the

choice of the length of the rolling estimation window can have profound effects on the performance of HAR models, with larger training windows generally leading to lower prediction errors.[9]

Information-set conventions: All forecasts are strictly real-time. At the forecast origin t, only information available through the close of day t is used. Monthly stock-level characteristics are lagged by at least one month to ensure observability; that is, the characteristic value assigned to day t is the most recent end-of-month observation available before t. The universe of the 500 largest stocks is reconstituted at the beginning of each annual re-estimation window based on market capitalization observed strictly before the start of the out-of-sample forecast period for that window. The stock-specific mean $\overline{RV}_i$ used for demeaning is computed exclusively from the rolling estimation window ending at day t, and is updated at each annual re-estimation date and held fixed throughout the subsequent forecast year.

Models: The models implemented in the paper are the HAR-Forest with minimum leaf size of 100 observations and all the variables summarized in Table 1 used as the splitting variables (HAR-Forest-Full) and its log-transformed version (Log-HAR-Forest-Full), the HAR-Forest with minimum leaf size of 100 observations and only RV^d, RV^w, and RV^m used as the splitting variables (HAR-Forest-RV), the HAR-Tree model of Corsi et al. (2012) and its logarithm version, conventional random forests with a minimum leaf size of five observations (Panel-RF), full panel HAR model (Panel-HAR), individual HARQ models, individual SHAR models, individual HExp models, individual HAR models (HAR) and their log-transformed version (Log-HAR). Finally, we follow Liu et al. (2015) and Li and Tang (2025) and implement an equally-weighted ensemble of all model forecasts. Table 2 lists all models considered in the paper.[10] The benchmark stock-specific HAR models, HARQ models, and the Panel-HAR model by Bollerslev et al. (2018) are estimated via ordinary-least-squares regression. For the SHAR model, we follow Patton and Sheppard (2015) and employ a weighted-least-squares (WLS) regression approach. The tree-structured HAR model proposed by Corsi et al. (2012), which employs a single-tree HAR framework for individual stocks,

[9]Inspired by these findings, in Section IA.6 of the Internet Appendix, we present robustness checks that explore variations in the length of the rolling estimation window.

[10]We discuss in Internet Appendix IA.8 the forecast results obtained with alternative ML-inspired benchmark models.

is implemented using the 1-day lagged realized volatility (RV) and 1-day lagged return as splitting variables. Finally, for the classical regression forest benchmark model, we choose to include 500 trees with a minimum node size of five and a feature subsampling rate of 1/3 which are the default values for random forest regression in many software implementations.

Sanity filter: None of the HAR models estimated guarantee non-negative RV predictions. For that purpose, we sanitize all predictions per stock i by replacing negative values with the empirical minimum value of the rolling window in-sample target variable of stock i in order to avoid negative forecasts. A similar filter is applied in Bollerslev et al. (2016), who instead replace negative predictions with the rolling unconditional average of the target variable. All models at the 22-day-ahead horizon have a filter-invocation rate of at most 0.2% of predictions.

Computational cost: The estimation of our model can be carried out on standard hardware and is highly parallelizable. Fitting a single tree on one core of an Apple Silicon M3 processor takes on average 44 seconds when employing five years of our cross-sectional data with the percentile grid $\{0.05, 0.06, \ldots, 0.95\}$, a minimum node size of 500 (versus the default of 100 used in the main analysis; estimation under the default setting is correspondingly slower), 18 splitting variables, and a feature subsampling rate of 1/3. The computational cost scales approximately linearly in the number of trees and in the number of grid points, and is highly parallelizable across trees. Hence, estimating a forest with 50 trees takes 37 minutes without making use of any parallelization to fit multiple trees at the same time. One easy way to reduce computational cost without large effects on predictive accuracy is to implement a coarser grid in the splitting optimization. Looking at a grid like {0.05, 0.10, ..., 0.90, 0.95} reduces the average estimation time of a single tree from 44 to 16 seconds. An R package for estimating the HAR-Forest is made available by the authors.[11]

3.3 Time-varying parameters

Unlike less-interpretable ML techniques such as neural networks, our local linear forest allows us to depict the state-implied time-varying HAR-Forest coefficients. At each point in time and for each stock, we traverse the trees in the forest to obtain the state-dependent coefficients as functions

[11]Note: URL to R package removed for paper submission.

of $Z_{\cdot,i,t}$. Figure 4 presents the bands containing 50%, 80%, and 95% of state-implied coefficients for the full sample across all stocks. The figure comprises four plots depicting the time-series of the three state-dependent HAR coefficients (β^d, β^w, β^m), and their monthly average estimated on the entire data set ($(\beta^d + \beta^w + \beta^m)/3$). The Figure is based on the HAR-Forest-Full model with a minimum leaf size of 100 observations, targeting the 22-day-ahead forecast horizon. The horizontal lines depict the coefficients implied by a time-invariant pooled HAR model.

Figure 4 reveals that the Panel-HAR estimates of β^d and β^w lie well below the interquartile range of the HAR-Forest's state-implied coefficients, indicating a severe downward bias when simple pooling is employed. In the first three panels, we observe substantial variation both over time and across stocks in $\beta^{\cdot}$. The variation across time and stocks is not limited to times of crises but can be observed across the entire sample. The fourth panel displays the variation in the average coefficient across lags. This reveals whether the individual coefficients merely offset one another or whether genuine variation exists in the persistence of the state-implied HAR models. As substantial time variation persists in this lag-averaged coefficient, we conclude that the HAR-Forest captures state-dependent persistence in the panel of stocks.

3.4 Accumulated local effects

In the following, we present ALE plots for the HAR-Forest-Full for 22-day-ahead forecasts on the full sample. We calculate the ALE plots per stock and aggregate the 1,131 stock-specific curves into fan charts. Figure 5 illustrates the ALE curves for the square root of RV^w, the VIX, and momentum.

The first row of Figure 5 plots the ALEs for realized volatility measured by $\sqrt{\text{RV}^w}$. We see that higher realized volatility is associated with small increases in β^d and β^m. In contrast, we see a pronounced negative effect in β^w as realized volatility increases. This pattern suggests that very high values of RV^w should not be overweighted for mid-term forecasts of volatility. The second row of Figure 5 displays the local effects of changes in VIX. Higher expected market volatility is associated with increases the values of β^d and β^w, while β^m decreases. This is consistent with empirical findings from Audrino and Knaus (2016), who document a greater importance of short-term volatility

components during turbulent periods. Finally, the bottom panel in Figure 5 examines the effect of momentum on the coefficients. The local effect of momentum is qualitatively different from those of realized volatility and VIX. The local effects are close to zero for all positive values of momentum. However, for negative momentum we see contrasting behavior. β^d and β^m show an upward trend with increasing momentum below 0, while β^w declines.

Figure 5 reveals heterogeneous local effects across covariates and HAR-Forest coefficients. The local effect of a covariate on one coefficient can be strong—such as the impact of VIX on β^m—while the same covariate may have a more moderate influence on other coefficients. Conversely, the realized volatility effect on β^d is not as pronounced as its effect on β^w.

As a measure of variable importance (VI), which is directly linked to the ALE plots, we calculate the standard deviation of the ALEs per stock $VI_{j,i}^{\beta^\cdot} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left[\widehat{ALE}_{j,i}^{\beta^\cdot}(z_{j,i,t})\right]^2}$, where T is the total number of daily observations. Figure 6 reports the average VI across stocks where the rows report the VIs per coefficient. We can see that the VIX, idiosyncratic volatility, and momentum have large effects on the three coefficients. RV^w and RV^m account for much of the variation in β^w and β^m. RV^w is more important than RV^d for both coefficients. Notably, the variable importance of realized kurtosis (RKurt) is relatively low, despite its documented effectiveness in improving HAR forecasts (Bollerslev et al., 2016). We attribute this to a signal quality issue: using RKurt as a proxy for measurement error in RV amounts to 'measuring noise with noise,' as RKurt is itself a high-variance statistic derived from a noisy proxy. Intuitively, on days where RV is already measured imprecisely, the fourth-moment estimator underlying RKurt is even more sensitive to outlying intraday returns. Our time-varying HAR-Forest coefficients depend less on RKurt but more on the level of the HAR variables. One expects that high values of RKurt tend to coincide with high values of RV^d, and high measurement error in RV^d is associated with even greater measurement error in RKurt.

The overall takeaway of this VI analysis is that our HAR-Forest-Full exploits a broad set of predictors, but the HAR regressors remain the dominant drivers. This aligns with our forecast evaluation in which we see gains in forecast performance by including covariates beyond RV^d, RV^w, and RV^m, but restricting the splitting variables to these three HAR regressors alone already

produces substantial outperformance relative to the benchmark.

3.5 Forecast evaluation metrics

In the remainder of this analysis, we follow Patton (2011) and use the SE and QLIKE loss. For a forecast of the average volatility over next h days $\widehat{RV}_{t+1:t+h|t}$ and its realization $RV_{t+1:t+h}$, the SE and QLIKE are defined as

$$\begin{aligned}
\text{SE}\left(RV_{t+1:t+h}, \widehat{RV}_{t+1:t+h}\right) &= \left(RV_{t+1:t+h} - \widehat{RV}_{t+1:t+h}\right)^2, \\
\text{QLIKE}\left(RV_{t+1:t+h}, \widehat{RV}_{t+1:t+h}\right) &= RV_{t+1:t+h}/\widehat{RV}_{t+1:t+h} - \ln\left(RV_{t+1:t+h}/\widehat{RV}_{t+1:t+h}\right) - 1.
\end{aligned}$$

As discussed in Patton (2011), the QLIKE is less sensitive to extreme observations than SE loss, but we report both for completeness.[12]

We consider the following forecasting schemes. Based on the information available on day t, cumulative volatility forecasts are computed for horizons of 1, 5, 22, 44, and 66 days. Forecast evaluation is based on the volatility proxy $RV_{i,t+1:t+h}$ and the respective forecast of model j for stock i is denoted by $\widehat{RV^j}_{i,t+1:t+h|t}$. For each loss function $\mathcal{L}$, we can measure the average loss of model j for stock i across time as

$$L_i^j = \frac{1}{|OOS|} \sum_{t \in OOS} \mathcal{L}(RV_{i,t+1:t+h}, \widehat{RV^j}_{i,t+1:t+h|t}), \tag{7}$$

where $|OOS|$ denotes the number of days in the OOS period. We denote the stock-specific loss of the benchmark forecast (i.e., the individual HAR model) by L_i^B. Finally, as a measure for the forecast accuracy of a particular model j relative to the benchmark, we consider the median loss ratio,

$$MedL^j = \underset{i=1,\ldots,N}{\text{median}} \frac{L_i^j}{L_i^B}. \tag{8}$$

[12] Another strand of literature applies ML beyond point forecasting. Luong and Dokuchaev (2018) apply random forest models to forecast future paths of realized volatility. Jiang et al. (2023) introduce a novel methodology that extracts trading signals from visual representations of price data, leveraging image-based inputs. A similar methodological approach is found in Kelly et al. (2023).

We also formally test for superior predictive ability. We base our analysis on the MCS approach introduced by Hansen et al. (2011), which is obtained as follows: Denote the set of all competing models by $\mathcal{M}$. For each loss function $\mathcal{L}$ and stock i, we define

$$d^{j,l}_{i,t+1:t+h} = \mathcal{L}\Big(RV_{i,t+1:t+h}, \widehat{RV}^{j}_{i,t+1:t+h|t}\Big) - \mathcal{L}\Big(RV_{i,t+1:t+h}, \widehat{RV}^{l}_{i,t+1:t+h|t}\Big)$$

as the difference in the respective loss of models j and l for the cumulative forecast horizon h. We compute the average loss difference per stock, $\bar{d}^{j,l}_i$, and calculate the test statistic

$$t^{j,l}_i = \frac{\bar{d}^{j,l}_i}{\sqrt{\widehat{\mathrm{Var}}(\bar{d}^{j,l}_i)}} \quad \text{for all } j,l \in \mathcal{M},$$

where $\widehat{\mathrm{Var}}(\bar{d}^{j,l}_i)$ denotes the estimate of the variance of the sample mean loss differential. The MCS test statistic is given by $T_{i,\mathcal{M}} = \max_{j,l\in\mathcal{M}} |t^{j,l}_i|$ and has the null hypothesis that all models have the same expected loss. Under the alternative, there is some model j that has an expected loss greater than the expected loss of all other models $l \in \mathcal{M} \setminus j$ for stock i. If the null hypothesis is rejected, the worst-performing model is removed from the set of models under consideration. The test is performed iteratively, until no further model can be eliminated. We denote the final set of surviving models for stock i by $\mathcal{M}_{i,MCS}$. This final set contains the best forecasting model with confidence level $1-\nu$. We set $\nu = 0.1$. This choice is common practice in the literature; see, for example, Laurent et al. (2013) and Liu et al. (2015). Since the asymptotic distribution of the test statistic $T_{i,\mathcal{M}}$ is nonstandard, we approximate it by block-bootstrapping as proposed by Hansen et al. (2011). In our analysis, 3,000 bootstrap replications at each stage were sufficient to obtain stable results.[13] As an aggregate measure, we calculate the share of stocks for which model j is in the model confidence set by

$$MCSR^j = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{j\in\mathcal{M}_{i,MCS}}.$$

Finally, we examine the performance of all models in two economic applications. For the

[13]For implementing the MCS procedure, we use the R package *rugarch* (Ghalanos, 2022) which includes the implementation used in the MFE Matlab toolbox by Kevin Sheppard. See: https://www.kevinsheppard.com/MFE_Toolbox.

cross-sectional forecast evaluation in Section 3.6, we implement a volatility management utility analysis per stock. A minimum-variance portfolio application follows in Section 3.7. Regarding asset-specific volatility timing, Bollerslev et al. (2018) consider an investor who possesses wealth W_t and invests x_t of his/her wealth to a risky asset with return $r_{t+1:t+h}$ and $(1 - x_t)$ to risk-free asset. The utility generated by such an allocation is given by

$$U(x_t) = W_t \left(x_t \mathrm{E}_t(r^e_{t+1:t+h}) - \frac{\gamma}{2} x_t^2 \mathrm{E}_t(RV_{t+1:t+h}) \right),$$

where $r^e_{t+1:t+h} = r_{t+1:t+h} - r^f_t$ is the excess return compared to return of the risk-free asset r^f_t and γ is the degree of relative risk aversion. Following Bollerslev et al. (2018), we assume that the Sharpe ratio (SR) of the risky asset is constant over time. This assumption, while strong, is standard in the volatility timing literature and is motivated by the desire to evaluate variance forecasts in isolation from expected-return estimation, which is notoriously noisy. Under this assumption, the utility can be expressed purely in terms of the volatility forecast:

$$U(x_t) = W_t \left(x_t SR \sqrt{\mathrm{E}_t(RV_{t+1:t+h})} - \frac{\gamma}{2} x_t^2 \mathrm{E}_t(RV_{t+1:t+h}) \right). \tag{9}$$

Utility (9) is maximized when $x^*_t = \frac{SR/\gamma}{\sqrt{\mathrm{E}_t(RV_{t+1:t+h})}}$. Correspondingly, the allocation of wealth to the risky asset depends on the ratio of the risk target to expected volatility. We consider the expected utility per unit of wealth for a mean-variance investor with an annualized SR of 0.4, a risk aversion parameter $\gamma = 2$, and an annualized volatility target of 20%. Now, we compute the average expected utility per unit of wealth $U_t(x^*_t)/W_t$ for every asset i and for every model j:

$$UoW^j_i = \frac{1}{|OOS|} \sum_{t \in OOS} \left(8\% \frac{\sqrt{RV_{i,t+1:t+h}}}{\sqrt{\widehat{RV^j}_{i,t+1:t+h|t}}} - 4\% \frac{RV_{i,t+1:t+h}}{\widehat{RV^j}_{i,t+1:t+h|t}} \right).$$

The average is taken over all OOS observations for every stock i and every model j. MedU and MCSU are defined analogously to MedL and MCSR.

3.6 Cross-sectional forecast evaluation results

Table 3 presents a comprehensive overview of the performance of our forecasting models during the OOS period 2005–2021. The Table is divided into several panels, each of which reports results for a different forecasting horizon. Panel A presents the results for one-day-ahead forecasts, Panel B for five-day-ahead forecasts, and Panel C for 22-day-ahead forecasts.[14] The models evaluated are listed in Table 2. As MedL is reported relative to the stock-individual HAR model, we only report the MCSR and MCSU statistics for the individual HAR model.[15] The model confidence sets are based on a broader range of models than displayed in Table 3 because they already include the additional benchmark models discussed in the robustness checks of Section IA.8 of the Internet Appendix.

For the one-day-ahead forecast horizon, the Log-HAR-Forest-Full model is the strongest model under squared-error loss, with the lowest SE-MedL of 91.7% and the highest SE-MCSR of 88.1%. Under QLIKE, the equally weighted ensemble (*Equal-Avg*) attains the lowest QLIKE-MedL of 80.5%, while Log-HAR-Forest-Full delivers the highest QLIKE-MCSR at 68.7%, only marginally above the 68.5% of Equal-Avg. The equally weighted ensemble also dominates in economic terms at this horizon, with the highest MedU of 105.3% and the highest MCSU inclusion rate of 73.7%. Thus, while the HAR-Forest specifications dominate the statistical loss criteria overall, the simple equal-weighted ensemble is particularly competitive at the short horizon and performs best for QLIKE and utility. The HExp model also performs comparatively well under QLIKE and utility. Among the remaining benchmark models, SHAR and Log-HAR-Tree are the strongest under SE loss, with SE-MCSR values of 54.2% and 47.4%, respectively, whereas Panel-HAR, HARQ, HAR, and the conventional panel random forest remain clearly behind the leading forest-based specifications.

The results for 5-day-ahead forecasts in Panel B reinforce the strong performance of the HAR-Forest class under statistical loss criteria. The Log-HAR-Forest-Full model attains the lowest SE-MedL and QLIKE-MedL, at 82.9% and 78.1%, respectively, and also records the highest MCS inclusion rates for both loss functions, 93.4% for SE and 90.6% for QLIKE. In utility terms, however,

[14] Additional results for 44-day-ahead and 66-day-ahead forecasts are reported in Section IA.5 of the Internet Appendix.

[15] The cross-sectional averaging of inclusion rates should be interpreted as a descriptive summary rather than a joint test of superior predictive ability across the entire panel. The portfolio-level results in Section 3.7 provide a complementary joint evaluation that implicitly accounts for cross-stock dependence.

HAR-Forest-Full slightly outperforms the other models, with MedU of 103.1% and MCSU of 81.7%, followed very closely by Log-HAR-Forest-Full with MedU of 102.6% and MCSU of 81.2%. The equally weighted ensemble remains competitive at this horizon as well, posting SE-MedL of 87.2%, QLIKE-MedL of 82.8%, MedU of 102.9%, and MCSU of 74.7%, but it no longer leads the table as it does at the one-day horizon.

The 22-day-ahead results in Panel C show a similar overall pattern, but with some differentiation across members of the HAR-Forest family. Log-HAR-Forest-Full remains the best model under SE loss, with the lowest SE-MedL of 84.2% and the highest SE-MCSR of 92.6%. Under QLIKE, however, the best performer is Log-HAR-Forest-RV, which attains the lowest QLIKE-MedL of 77.9% and the highest QLIKE-MCSR of 87.2%. In utility terms, HAR-Forest-RV yields the highest MedU of 103.8%, while Log-HAR-Forest-RV achieves the highest MCSU inclusion rate at 87.2%. Hence, at the monthly horizon, the best-performing models continue to be forest-based, but the leadership is split across the Full and RV variants depending on the evaluation criterion.

Taken together, the evidence in Table 3 shows that the HAR-Forest models, especially the log-transformed variants, deliver the most robust improvements in cross-sectional forecasting performance across horizons and loss functions. At the same time, the results also highlight that simple forecast combination can be very effective: the equally weighted ensemble is the best-performing specification for one-day-ahead QLIKE loss and utility, and remains highly competitive at longer horizons.

To put the increases in utility into perspective, we examine the stock identified as the cross-sectional median of utility ratios, which in our sample is Eastman Chemical Company. In untabulated results, we find that the difference in utility for this stock is equal to 9.2 basis points. To quantify the economic gains of a 9.2-basis-point increase, we provide an example along the lines of Bollerslev et al. (2018): when investing \$10 million in Eastman Chemical Company, a fund would be willing to pay \$9,200 to switch from individual HAR models to our HAR-Forest-Full specification. Summing the stock-level utility differences across all 1,131 stocks in the cross-section yields a cumulative gain of 279 basis points from HAR-Forest-Full relative to the individual HAR model. For a fund investing \$10 million in each of the 1,131 stocks, this cumulative gain corresponds to a willingness

to pay of approximately $279,000 to switch to our HAR-Forest specification.[16]

For further comparison of the models across different volatility regimes, we disaggregate the SE-MedL and QLIKE-MedL losses across different deciles of volatility for 22-day-ahead forecasts. The results are reported in Figure 7 and, again, illustrate the superior performance of HAR-Forest-Full and Log-HAR-Forest-Full relative to the benchmark models across volatility regimes. We also observe that the conventional random forest's SE-MedL and QLIKE-MedL rise sharply in the higher-volatility deciles, rendering this model unreliable in such regimes. We will see in Section 3.7 that the improved univariate forecast performance translates into superior minimum-variance portfolio performance, and in Internet Appendix IA.4 that these gains are broadly consistent across firm characteristics.

3.7 From statistical accuracy to economic value

We next evaluate whether the improved forecasts translate into better portfolio performance by constructing global minimum-variance (GMV) portfolios under each forecasting model. This criterion is particularly attractive for our application because its optimal weights depend exclusively on the conditional covariance matrix and are invariant to expected-return estimates, whose estimation error often overwhelms any gains from improved risk forecasts.

To obtain the GMV, we compute the time-t daily conditional covariance matrix associated with forecast scheme k using the standard factorization

$$\widehat{\mathbf{\Sigma}}_t^{(k)} \;=\; \mathbf{D}_t^{(k)}\, \widetilde{\mathbf{R}}_t\, \mathbf{D}_t^{(k)}, \tag{10}$$

where $\mathbf{D}_t^{(k)} = \operatorname{diag}\left(\sqrt{\widehat{RV}_{1,t}^{(k)}}, \ldots, \sqrt{\widehat{RV}_{N_t,t}^{(k)}}\right)$ and $\widetilde{\mathbf{R}}_t$ is the correlation matrix and $\widetilde{\mathbf{R}}_t$ is the correlation matrix which we estimate using our 5-minute intraday data. We employ the same $\widetilde{\mathbf{R}}_t$ across all models such that any difference in portfolio performance is solely driven by the realized variance forecasts.

[16] In Section IA.6 we provide evidence that the results presented in Table 3 are largely robust to the length of the estimation window. We also present additional results for other choices of hyperparameters, namely a smaller number of trees in the forest and a larger minimum node size.

We compute daily realized correlation matrices from 5-minute log-returns and average the most recent 22 daily matrices to reduce estimation noise.[17] All models use the same correlation matrix so that portfolio performance differences are driven solely by variance forecasts.

We consider an investor with a 22-day investment horizon. Consistent with this notion, we populate $\mathbf{D}_t^{(k)}$ with 22-day-ahead forecasts of the daily realized variances obtained with the different model specifications.[18] Finally, the GMV portfolio is computed as

$$\mathbf{w}_t^{\mathrm{GMV},(k)} = \frac{\widehat{\mathbf{\Sigma}}_t^{(k)\,-1}\mathbf{1}}{\mathbf{1}^\top \widehat{\mathbf{\Sigma}}_t^{(k)\,-1}\mathbf{1}}, \tag{11}$$

where $\mathbf{1}$ is an N_t-dimensional vector of ones. The realized GMV portfolio return over the h-day holding period is $r_{t+1:t+h}^{\mathrm{GMV}} = \mathbf{w}_t^\top \mathbf{r}_{t+1:t+h}$, where $\mathbf{w}_t$ is the vector of GMV weights chosen at time t and $\mathbf{r}_{t+1:t+h}$ is the vector of individual asset returns realized over the holding period $[t, t+h]$.

Table 4 reports performance statistics for the GMV portfolios. In addition to the model specifications listed in Table 3, we follow Li and Tang (2025) and include results obtained from a "perfect risk model." For this, the diagonal elements of the conditional variance matrix $\mathbf{D}_t^{(k)}$ are populated by ex post observed (instead of forecast) realized variance estimates. All reported statistics are based on out-of-sample observations.[19]

As anticipated, the perfect risk model delivers the best overall performance. Among the forecasting models evaluated, the HAR-Forest-Full and HAR-Forest-RV specifications achieve the lowest portfolio return standard deviations among all forecasting models (3.40% and 3.38%, respectively), both statistically lower than the individual HAR model (4.15%) at the 1% level, with Sharpe ratio improvements significant at the 5% level. The Log-HAR-Forest-Full achieves a monthly standard deviation of 2.97%, which is likewise significantly lower than the HAR benchmark,

[17] Using 5-minute returns balances intraday information and synchronicity; see Barndorff-Nielsen et al. (2009). Aggregating over 22 days reduces the impact of daily estimation error. For discussion of asynchronous trading bias, see Epps (1979) and Liu (2009).

[18] We also considered 1-day and 5-day investment horizons. The results are qualitatively similar to those based on 22-day.

[19] Because we have overlapping returns due to a daily forecast scheme combined with a 22-day investment period, we report the average of 22 portfolio investments: An investor who rebalances on day 22, 44, 66, etc. of the out-of-sample period, an investor who rebalances on day 23, 45, 67, etc., ..., and an investor who rebalances on day 43, 65, 87, etc. such that each individual investor has non-overlapping returns. This is especially important for the maximum drawdown measure.

and produces portfolios with a smaller maximum drawdown (−30.1% vs. −35.3%). However, the Sharpe ratio improvements for the log-transformed specifications are not statistically significant at conventional levels. Taken together, the results in Table 4 suggest that the enhanced forecasting accuracy of our Log-HAR-Forest specification translates into improved economic performance, as reflected in GMV portfolios with substantially lower risk in comparison to all competing specifications.

Table 5 reports descriptive statistics for the GMV portfolio weights generated under each forecasting model. We find that the perfect risk model exhibits the highest degree of concentration, as indicated by the highest maximum weight (0.281) and largest weight dispersion (standard deviation of 0.018). In contrast, all forecast-based models display lower dispersion and smaller maximum portfolio weights. Among the evaluated forecasting specifications, the Log-HAR-Forest models exhibit slightly higher maximum allocations (up to 0.136) relative to their HAR-Forest counterparts, suggesting that the log-specification produces more differentiated risk estimates across stocks. Overall, these statistics suggest that enhanced predictive accuracy, as documented previously, also manifests itself through more confident – and economically meaningful – portfolio allocations.

3.8 Robustness checks

We now examine the sensitivity of the results reported in Section 3.6 along five dimensions: firm characteristics, forecast horizon, estimation window length, hyperparameter specification, and the choice of ML benchmarks. In each case, the relative ranking of models is preserved and the outperformance of the HAR-Forest is maintained.

Firm characteristics. Internet Appendix IA.4 investigates whether forecast improvements are concentrated among stocks with particular firm characteristics. For each stock, we compute Diebold–Mariano (DM) test statistics comparing the HAR-Forest-Full against individually estimated HAR models and regress these statistics on characteristics drawn from Gu et al. (2020), including book-to-market ratios, cash holdings, illiquidity, leverage, size, and trading volume. The intercept is positive

and statistically significant with t-statistics exceeding 5 across all specifications, indicating broad outperformance of the HAR-Forest-Full. When characteristics are introduced individually, only cash holdings exhibit a statistically significant association with the DM statistics at the 1% level, with an R^2 of 1.34%. A joint regression including all characteristics simultaneously yields an adjusted R^2 below 2.1%. These conclusions hold across two non-overlapping subsamples—2005–2013M6 and 2013M7–2021—where the intercept remains significant with t-statistics above 5.9 and the maximum R^2 does not exceed 2.49%. The forecasting gains of the HAR-Forest-Full are therefore not attributable to any identifiable subset of firm characteristics.

Longer forecast horizons. Internet Appendix IA.5 extends the evaluation to 44- and 66-day-ahead horizons. The Log-HAR-Forest-Full and Log-HAR-Forest-RV models achieve the lowest or near-lowest MedL under both the SE and QLIKE criteria at both horizons, while maintaining competitive utility gains. The relative ranking of models is preserved throughout, indicating that the outperformance documented in Section 3.6 does not diminish at longer horizons.

Estimation window length. Chassot and Audrino (2025) document that the length of the rolling estimation window materially affects HAR model performance, with longer windows generally yielding lower prediction errors. Internet Appendix IA.6 presents results for three- and ten-year estimation windows, holding the out-of-sample period fixed at 2005–2021. The HAR-Forest-Full and Log-HAR-Forest-Full models outperform competing specifications under both window lengths.

Hyperparameter sensitivity. Internet Appendix IA.7 assesses sensitivity to three key hyperparameter choices: the number of trees (200 vs. 50, Section IA.7.1), the minimum node size (100 vs. 500, Section IA.7.2), and the feature subsampling rate (1/3 vs. 1/2, Section IA.7.3). Across all configurations and forecast horizons, the relative ranking of models is unchanged. These findings are consistent with the well-documented robustness of random forest-type estimators to hyperparameter specification (Breiman, 2001).

Alternative machine-learning benchmarks. Internet Appendix IA.8 evaluates two additional classes of ML-inspired benchmarks: models incorporating cross-sectional volatility spillovers and models that relax the linearity of the HAR framework via gradient boosting, random forests, and neural networks applied to HAR regressors. None of these alternatives consistently outperforms the standard HAR model under either the SE or QLIKE loss, and all fall substantially short of the HAR-Forest across statistical and economic criteria.

4 Concluding remarks

We combine the ML technique of random forests with the well-established HAR model to forecast realized volatilities in a large cross-section of individual stocks. We propose to model the coefficients of the HAR model as nonparametric functions of state variables (e.g., realized measures, the VIX, stock market betas, or momentum). We achieve this by employing these characteristics as splitting variables in decision trees. We address data sparsity issues in deeper trees by estimating panel HAR models in each terminal node.

In an empirical application using the 500 largest stocks from NYSE/NASDAQ/AMEX dynamically selected to avoid survivorship bias from 2000 until 2021, we document that our HAR-Forest model achieves statistically significantly lower QLIKE and SE losses and higher utility gains when compared to the benchmark models across different forecast horizons. Similarly, the improved forecast performance of the HAR-Forest leads to minimum-variance portfolio returns with lower volatility and higher Sharpe ratios when compared to standard volatility models.

Exploiting the interpretability of our hybrid approach, we find that the time-varying HAR-Forest parameters differ substantially from the fixed coefficients of the Panel-HAR. Based on our accumulated local effects analysis, we find that short-term volatility components have greater importance during turbulent periods, such as when the VIX is above 20. Our variable importance analysis shows that VIX, idiosyncratic volatility, momentum, and weekly and monthly realized volatility are the main additional drivers of the time-varying parameters. Finally, the forecasting gains are not confined to a specific subset of stocks but hold broadly across firm characteristics and

volatility regimes.

References

Aït-Sahalia, Y. and D. Xiu (2019). A Hausman test for the presence of market microstructure noise in high frequency data. *Journal of Econometrics 211*(1), 176–205.

Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez (2015). Does realized skewness predict the cross-section of equity returns? *Journal of Financial Economics 118*(1), 135–167.

Andersen, T. G. and T. Bollerslev (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review 39*(4), 885–905.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of Financial Economics 61*(1), 43–76.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2011). Realized volatility forecasting and market microstructure noise. *Journal of Econometrics 160*(1), 220–234.

Andersen, T. G., D. Dobrev, and E. Schaumburg (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics 169*(1), 75–93.

Apley, D. W. and J. Zhu (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(4), 1059–1086.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Annals of Statistics 47*(2), 1179–1203.

Audrino, F., C. Huang, and O. Okhrin (2019). Flexible har model for realized volatility. *Studies in Nonlinear Dynamics & Econometrics 23*(3).

Audrino, F. and S. D. Knaus (2016). Lassoing the HAR Model: A model selection perspective on realized volatility dynamics. *Econometric Reviews 35*, 1485–1521.

Bali, T. G., A. Goyal, D. Huang, F. Jiang, and Q. Wen (2020). Predicting corporate bond returns: Merton meets machine learning. *Georgetown McDonough School of Business Research Paper* (3686164), 20–110.

Barigozzi, M., C. Brownlees, G. M. Gallo, and D. Veredas (2014). Disentangling systematic and idiosyncratic dynamics in panels of volatility measures. *Journal of Econometrics 182*(2), 364–384.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2009). Realized kernels in practice: Trades and quotes. *Econometrics Journal 12*(3).

Barndorff-Nielsen, O. E., S. Kinnebrock, and N. Shephard (2010). Measuring downside risk—realized semivariance. *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*, 1–22.

Bekaert, G. and M. Hoerova (2014). The VIX, the variance premium and stock market volatility. *Journal of Econometrics 183*(2), 181–190.

Bettencourt, L. O., A. Tetereva, and A. Petukhina (2024). Advancing markowitz: Asset allocation forest.

Bollerslev, T., B. Hood, J. Huss, and L. H. Pedersen (2018). Risk everywhere: Modeling and managing volatility. *Review of Financial Studies 31*(7), 2730–2773.

Bollerslev, T., S. Z. Li, and B. Zhao (2019). Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis 55*(3), 751–781.

Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2016). Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting. *Journal of Econometrics 192*(1), 1–18.

Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2022). Realized semibetas: Disentangling "good" and "bad" downside risks. *Journal of Financial Economics 144*(1), 227–246.

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

Brownlees, C. and A. B. Souza (2021). Backtesting global growth-at-risk. *Journal of Monetary Economics 118*(1), 312–330.

Brownlees, C. T. (2019). Hierarchical GARCH. *Journal of Empirical Finance 51*(December 2018), 17–27.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics 18*(3), 502–531.

Carr, P., L. Wu, and Z. Zhang (2020). Using machine learning to predict realized variance. *Journal of Investment Management 18*(2).

Chassot, J. and F. Audrino (2025). Hard to beat: The overlooked impact of rolling windows in the era of machine learning. *International Journal of Forecasting*. In press, corrected proof; available online 13 Aug 2025.

Christensen, K., M. Siggaard, and B. Veliyev (2023). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics 21*(5), 1680–1727.

Cipollini, F., G. M. Gallo, and E. Otranto (2021). Realized volatility forecasting: Robustness to measurement errors. *International Journal of Forecasting 37*(1), 44–57.

Cong, L., G. Feng, J. He, and X. He (2025). Growing panel trees to harvest basis portfolios and pricing kernels. *Journal of Financial Economics*, Forthcoming.

Cong, L. W., G. Feng, J. He, and J. Li (2023). Sparse modeling under grouped heterogeneity with an application to asset pricing. Technical report, National Bureau of Economic Research.

Cong, L. W., G. Feng, J. He, and Y. Wang (2024). Mosaics of predictability. *Available at SSRN 4853767*.

Conrad, C., O. Kleen, and R. Lönn (2025). Volatility forecasting for low-volatility investing. *International Journal of Forecasting*.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7*(2), 174–196.

Corsi, F., F. Audrino, and R. Renó (2012). Har modeling for realized volatility forecasting. In *Handbook of Volatility Models and Their Applications*, pp. 363–382. John Wiley & Sons.

Corsi, F., S. Mittnik, C. Pigorsch, and U. Pigorsch (2008). The volatility of realized volatility. *Econometric Reviews 27*(1-3), 46–78.

DeMiguel, V., J. Gil-Bazo, F. J. Nogales, and A. A. Santos (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics 150*(3), 103737.

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

Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association 74*(366), 291–298.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

Fisher, A., C. Rudin, and F. Dominici (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research 20*(177), 1–81.

Freire, G. and O. Kleen (2022). Equity options and firm characteristics. *Available at SSRN: 4342597*.

French, K. R. and R. Roll (1986). Stock return variances: The arrival of information and the reaction of traders*. *Journal of Financial Economics 17*(1), 5–26.

Friedberg, R., J. Tibshirani, S. Athey, and S. Wager (2020). Local linear forests. *Journal of Computational and Graphical Statistics 30*(2), 503–517.

Gallo, G. M. and E. Otranto (2015). Forecasting realized volatility with changing average levels. *International Journal of Forecasting 31*(3), 620–634.

Ghalanos, A. (2022). *rugarch: Univariate GARCH models*. R package version 1.4-7.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2005). There is a risk-return trade-off after all. *Journal of Financial Economics 76*(3), 509–548.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Goulet Coulombe, P. (2024). The macroeconomy as a random forest. *Journal of Applied Econometrics 39*(3), 401–421.

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

Guijarro-Ordonez, J., M. Pelger, and G. Zanotti (2025). Deep learning statistical arbitrage. *Management Science*. Published online December 4, 2025.

Hansen, P. R. and A. Lunde (2014). Estimating the persistence and the autocorrelation function of a time series that is measured with error. *Econometric Theory 30*(1), 60–93.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Harvey, C. R., E. Hoyle, R. Korgaonkar, S. Rattray, M. Sargaison, and O. Van Hemert (2018). The impact of volatility targeting. *Journal of Portfolio Management 45*(1), 14–33.

Jiang, J., B. Kelly, and D. Xiu (2023). (re-) imag (in) ing price trends. *The Journal of Finance 78*(6), 3193–3249.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics 150*(1), 94–138.

Kelly, B. T., B. Kuznetsov, S. Malamud, and T. A. Xu (2023). Deep learning from implied volatility surfaces. *Swiss Finance Institute Research Paper* (23-60).

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Laurent, S., J. V. K. Rombouts, and F. Violante (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics 173*(1), 1–10.

Li, B. and A. G. Rossi (2020). Selecting mutual funds from the stocks they hold: A machine learning approach. *Available at SSRN 3737667*.

Li, S. Z. and Y. Tang (2025). Automated volatility forecasting. *Management Science 71*(7), 6248–6274.

Liu, L. Y., A. J. Patton, and K. Sheppard (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics 187*(1), 293–311.

Liu, Q. (2009). On portfolio optimization: How and when do we benefit from high-frequency data? *Journal of Applied Econometrics 24*(4), 560–582.

Luong, C. and N. Dokuchaev (2018). Forecasting of realised volatility with the random forests algorithm. *Journal of Risk and Financial Management 11*(4), 61.

Mücher, C. (2021). Artificial neural network based non-linear transformation of high-frequency returns for volatility forecasting. *Frontiers in Artificial Intelligence 4*.

Müller, U. A., M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward (1993). Fractals and intrinsic time: A challenge to econometricians. *Unpublished manuscript, Olsen & Associates, Zürich*, 130.

Pakel, C., N. Shephard, and K. Sheppard (2011). Nuisance parameters, composite likelihoods and a panel of GARCH models. *Statistica Sinica 21*(1), 307–329.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics 160*(1), 246–256.

Patton, A. J. and K. Sheppard (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics 97*(3), 683–697.

Patton, A. J. and Y. Simsek (2023). Generalized autoregressive score trees and forests. *Available at arXiv: https://arxiv.org/abs/2305.18991*.

Pesaran, M. H. and A. Pick (2011). Forecast combination across estimation windows. *Journal of Business and Economic Statistics 29*(2), 307–318.

Pesaran, M. H. and R. Smith (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of econometrics 68*(1), 79–113.

Politis, D. and J. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Rahimikia, E. and S.-H. Poon (2020). Machine learning for realised volatility forecasting. *Available at SSRN 3707796*.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Tables

Table 1: Cross-sectional summary statistics

	Mean	Median	Std	Skew	Kurt	Min	Max
Panel A: Daily variables							
RV^d	9.044	3.694	31.272	19.914	715.667	0.240	1178.270
RV^w	9.051	4.159	20.547	9.477	155.805	0.489	391.670
RV^m	9.073	4.733	14.924	5.334	45.816	0.833	159.059
RV^+	3.407	1.535	8.449	13.283	371.629	0.077	256.804
RV^-	3.395	1.526	8.814	14.116	432.121	0.075	286.006
MedRV	5.647	2.658	12.889	11.375	277.786	0.138	360.654
Ret^d	0.033	0.040	2.706	-0.479	25.043	-30.849	24.830
Ret^w	0.033	0.058	1.329	-0.545	14.362	-11.357	9.432
Ret^m	0.032	0.063	0.630	-0.751	9.431	-4.167	3.175
RSkew	0.029	0.016	1.216	0.058	6.596	-7.162	7.298
RKurt	6.891	5.340	5.154	3.927	27.620	2.140	64.387
VIX	19.968	17.856	8.679	2.117	10.288	9.405	78.170
Panel B: Monthly variables							
Idiovol	0.046	0.042	0.016	0.655	2.604	0.027	0.080
Beta	1.117	1.107	0.354	0.095	2.623	0.480	1.858
Mom	0.171	0.110	0.445	1.021	6.553	-0.602	2.130
Size	17.211	14.598	9.981	0.852	3.918	4.694	48.426
Volume	16.137	16.235	0.803	-0.353	3.215	14.005	17.819
Panel C: Cross-sectional variable							
$\overline{RV}$	9.061	6.855	6.950	2.090	8.802	1.869	55.151

Notes: This table reports cross-sectional averages of stock-specific statistics of our splitting variables across all 1,131 stocks. A stock is included if it appears at least once among the 500 largest common stocks during the out-of-sample period. We report the cross-sectional average of the number of observations, mean, median, standard deviation, skewness, kurtosis, minimum, and maximum across stocks. The sample period is 2000–2021.

Table 2: Models considered in the paper

Model Acronym	**Description**
HAR-Forest-Full	HAR model with time-varying coefficients via random forests, using all available features listed in Table 1 as splitting variables
HAR-Forest-RV	HAR model with time-varying coefficients via random forests, using only HAR regressors as splitting variables
Log-HAR-Forest-Full	Same as HAR-Forest-Full but estimated on the log of realized volatility
Log-HAR-Forest-RV	Same as HAR-Forest-RV but estimated on the log of realized volatility
Panel-RF	Random forest model applied to the cross-section of stocks
HAR-Tree	Single-tree variant of the heterogeneous autoregressive model
Log-HAR-Tree	Same as HAR-tree but using log of realized volatility
Panel-HAR	Pooled panel version of the heterogeneous autoregressive model
HExp	Heterogeneous exponential realized volatility model
HARQ	HAR model incorporating realized quarticity to handle measurement error
SHAR	HAR model extended with realized semivariances of positive and negative returns
Log-HAR	Heterogeneous autoregressive model using log of realized volatility
HAR	Baseline heterogeneous autoregressive model
Ensemble	Equally-weighted average of the forecasts from all competing models listed above

Table 3: Unconditional forecast evaluation

	SE		QLIKE		Utility	
Model	MedL	MCSR	MedL	MCSR	MedU	MCSU
Panel A: 1-day-ahead						
HAR-Forest-Full	0.931	0.639	0.845	0.374	1.034	0.447
HAR-Forest-RV	0.943	0.581	0.865	0.263	1.033	0.445
Log-HAR-Forest-Full	**0.917**	**0.881**	0.835	**0.687**	1.018	0.387
Log-HAR-Forest-RV	0.924	0.768	0.847	0.469	1.014	0.302
Panel-RF	0.974	0.455	0.886	0.293	1.036	0.531
HAR-Tree	1.021	0.204	1.046	0.050	0.973	0.086
Log-HAR-Tree	0.959	0.474	0.888	0.244	1.006	0.170
Panel-HAR	1.021	0.129	1.052	0.017	0.992	0.057
HARQ	0.994	0.199	1.021	0.034	0.973	0.063
SHAR	0.942	0.542	0.880	0.227	1.024	0.265
Log-HAR	0.955	0.420	0.884	0.205	1.006	0.138
HExp	0.999	0.450	0.847	0.514	1.045	0.568
Equal-Avg	0.926	0.662	**0.805**	0.685	**1.053**	**0.737**
HAR	—	0.189	—	0.023	—	0.059
Panel B: 5-day-ahead						
HAR-Forest-Full	0.842	0.728	0.802	0.663	**1.031**	**0.817**
HAR-Forest-RV	0.866	0.636	0.824	0.523	1.030	0.742
Log-HAR-Forest-Full	**0.829**	**0.934**	**0.781**	**0.906**	1.026	0.812
Log-HAR-Forest-RV	0.849	0.796	0.794	0.721	1.025	0.719
Panel-RF	0.898	0.600	0.846	0.484	1.028	0.725
HAR-Tree	1.010	0.332	0.967	0.168	1.000	0.259
Log-HAR-Tree	0.921	0.633	0.852	0.453	1.015	0.438
Panel-HAR	1.024	0.314	1.068	0.097	0.990	0.221
HARQ	0.989	0.322	0.995	0.120	0.997	0.205
SHAR	0.913	0.542	0.881	0.345	1.018	0.451
Log-HAR	0.909	0.647	0.851	0.417	1.014	0.423
HExp	1.043	0.427	1.007	0.306	0.991	0.388
Equal-Avg	0.872	0.681	0.828	0.532	1.029	0.747
HAR	—	0.324	—	0.122	—	0.197
Panel C: 22-day-ahead						
HAR-Forest-Full	0.873	0.733	0.803	0.713	1.035	0.780
HAR-Forest-RV	0.905	0.660	0.788	0.724	**1.038**	0.782
Log-HAR-Forest-Full	**0.842**	**0.926**	0.782	0.868	1.031	0.866
Log-HAR-Forest-RV	0.859	0.853	**0.779**	**0.872**	1.032	**0.872**
Panel-RF	1.083	0.455	0.888	0.484	1.021	0.629
HAR-Tree	0.998	0.509	0.972	0.322	1.001	0.475
Log-HAR-Tree	0.923	0.791	0.853	0.673	1.018	0.727
Panel-HAR	1.008	0.555	1.042	0.247	0.994	0.406
HARQ	0.985	0.505	0.984	0.272	1.001	0.408
SHAR	0.930	0.649	0.892	0.497	1.017	0.595
Log-HAR	0.918	0.802	0.854	0.677	1.017	0.729
HExp	1.010	0.620	0.997	0.381	0.997	0.497
Equal-Avg	0.878	0.738	0.832	0.620	1.027	0.727
HAR	—	0.528	—	0.264	—	0.401

Notes: We report the median loss ratio (MedL) relative to the HAR benchmark model (lower is better) and the MCS inclusion rate (MCSR) across stocks (higher is better); see Section 3.5. The last two columns report similar figures for the utility framework by Bollerslev et al. (2018). We report the median utility ratio (MedU) relative to the individual HAR benchmark model (higher is better), and the MCS inclusion rate of the utility measure (MCSU) across stocks (higher is better). Results for individually fitted HAR models are found in the rows labeled "HAR". The number of trees per HAR-Forest is 200. The number of trees in the random forest is 500. The forecasts are issued daily and, hence, overlapping. The OOS period is 2005–2021. Averages are taken across 1,131 stocks. Figures in bold indicate the best outcome per panel and evaluation criterion (lowest for MedL, highest for MCSR, MedU, and MCSU). The model confidence sets are calculated including the additional benchmark models in Section IA.8.

Table 4: Minimum-variance portfolio performance

Model	**Mean ret. (%)**	**Sd. (%)**	**SR**	**FF3 alpha (%)**	**Drawdown (%)**
Perfect risk model	1.358	2.562***	0.534***	1.077***	-12.148
HAR-Forest-Full	1.074	3.400***	**0.317****	0.620***	-29.292
HAR-Forest-RV	1.062	3.378***	0.316**	0.611***	-30.645
Log-HAR-Forest-Full	0.893	**2.970*****	0.301	0.530***	-30.071
Log-HAR-Forest-RV	0.838	2.929***	0.287	0.478***	-30.265
Panel-RF	1.040	3.661***	0.285	0.523***	-31.657
HAR-Tree	1.034	4.465	0.236	0.474***	-41.836
Log-HAR-Tree	0.963	3.340***	0.290	0.511***	-32.165
Panel-HAR	1.073	3.529***	0.305*	0.598***	-29.945
HARQ	1.153	4.843	0.239	0.514***	-43.185
SHAR	1.087	4.254	0.257	0.513***	-37.850
Log-HAR	0.958	3.455***	0.278	0.508***	-36.435
HExp	1.074	4.009	0.272	0.524***	-35.650
Equal-Avg	1.022	3.680***	0.280	0.547***	-34.638
HAR	1.170	4.149	0.284	0.583***	-35.283

Notes: The table reports monthly mean portfolio returns, their standard deviations (Sd.), Sharpe ratios (SR), three-factor alphas (FF3 alphas) based on the Fama and French (1993) factor model, and maximum drawdown (Drawdown). All results are calculated using out-of-sample observations discussed in Section 3.6. One, two, and three asterisks denote statistical significance of the difference relative to the individual HAR model at the 10%, 5%, and 1% levels, respectively. For mean returns and FF3 alphas, we employ Newey-West standard error for the statistical inference. p-values for testing the portfolio volatilities and Sharpe ratios against the HAR model benchmark are computed using the stationary bootstrap of Politis and Romano (1994) with B=1,000 resamples and expected block size $b = \lceil 1.75 \times h^{1/3} \rceil = 5$ for h=22. As a robustness check, we verified that the significance of all reported results is preserved when using $b \in \{10, 22, 44\}$. The lowest portfolio standard deviation and the highest Sharpe ratio excluding the perfect risk model are highlighted in bold.

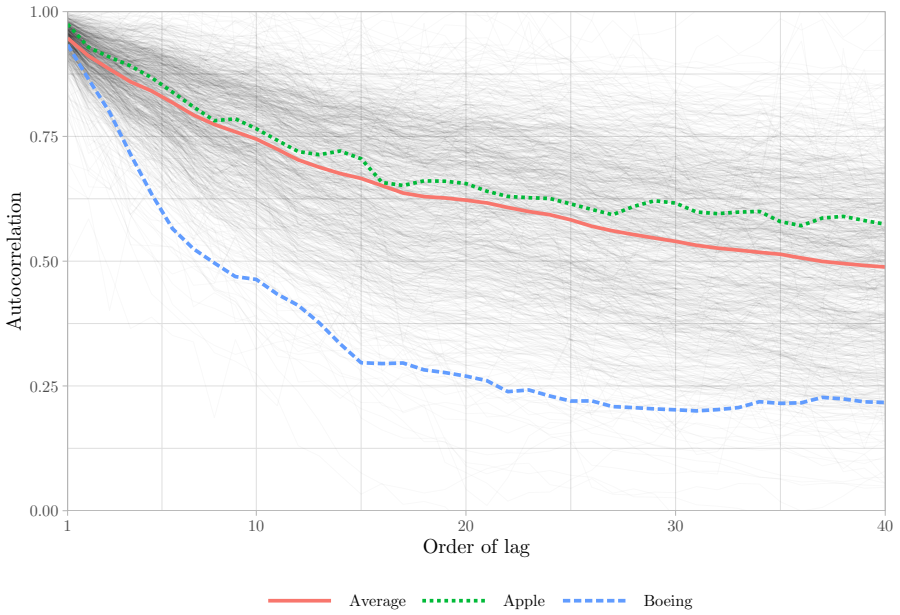
Table 5: Descriptive statistics of minimum-variance portfolio weights

Model	**Min. Weights**	**Sd. Weights**	**Max. Weights**
Perfect risk model	-0.034	0.018	0.281
HAR-Forest-Full	-0.053	0.016	0.081
HAR-Forest-RV	-0.052	0.015	0.085
Log-HAR-Forest-Full	-0.049	0.017	0.152
Log-HAR-Forest-RV	-0.046	0.018	0.197
Panel-RF	-0.052	0.015	0.080
HAR-Tree	-0.044	0.016	0.172
Log-HAR-Tree	-0.047	0.017	0.155
Panel-HAR	-0.053	0.015	0.077
HARQ	-0.049	0.015	0.117
SHAR	-0.051	0.015	0.094
Log-HAR	-0.048	0.017	0.163
HExp	-0.051	0.015	0.084
Equal-Avg	-0.053	0.016	0.088
HAR	-0.050	0.015	0.091

Notes: The table reports descriptive statistics of global minimum-variance portfolio weights (minimum, maximum, and standard deviation).

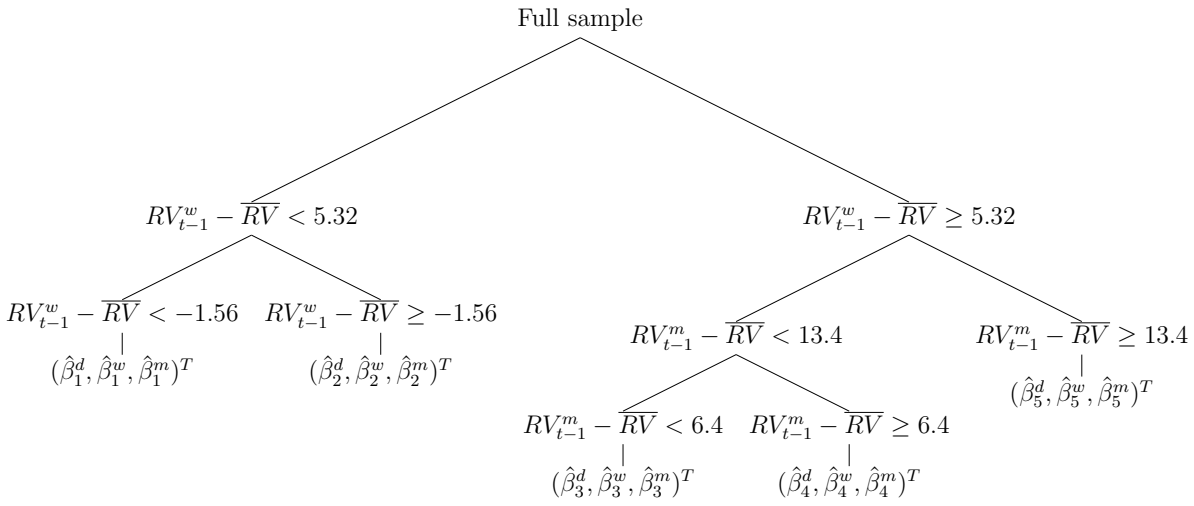
Figures

Figure 1: Autocorrelation function across stocks



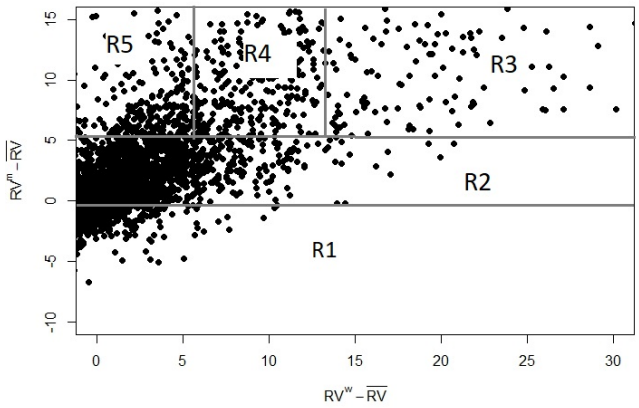
Notes: We depict the empirical autocorrelation function (ACF) for RV^d across the subset of stocks in our sample for which we have at least 2500 observations per stock (886 stocks). The ACF is estimated using the instrumental variables regression proposed by Hansen and Lunde (2014) with 4 to 10 lagged values of RV^d. ACF of Apple Inc. and Boeing are depicted in green (dotted) and blue (dashed) respectively. The cross-sectional average ACF is depicted in red (solid).

Figure 2: Example of a tree with HAR models in each leaf



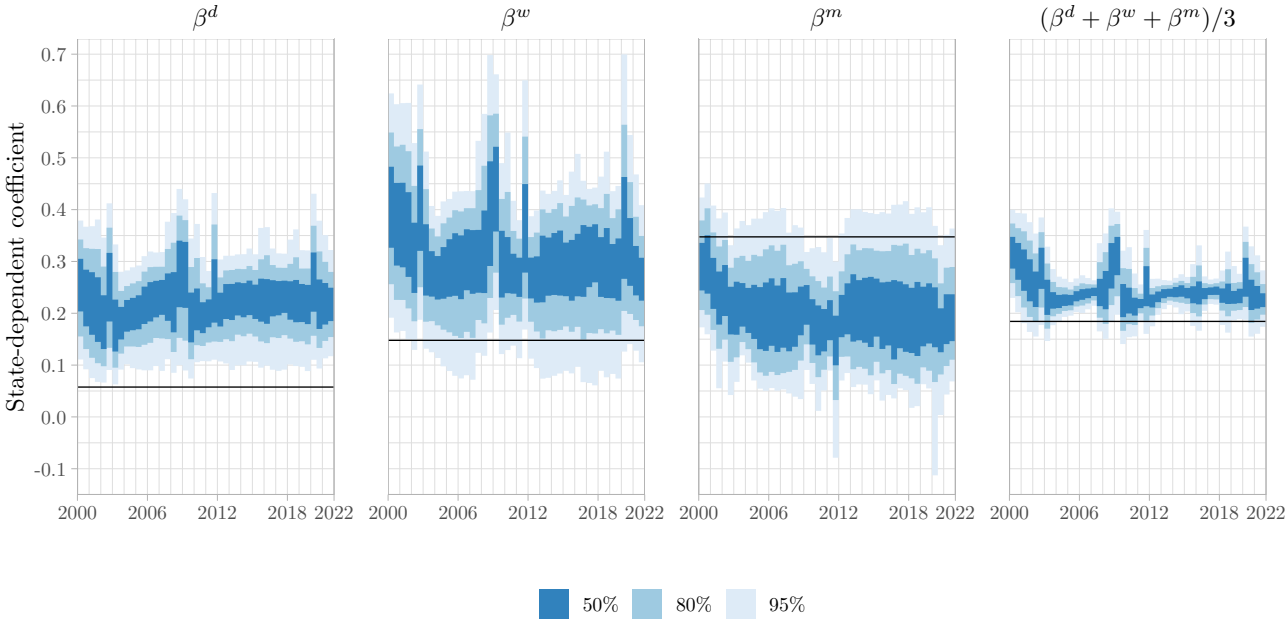
Notes: At each node, we decide how to go through the tree based on the value of a splitting variable, e.g., $RV^w - \overline{RV}$ at the parent node. Each leaf consists of a set of coefficients that constitute the local HAR model for this leaf.

Figure 3: Example of the number of observations per leaf from Figure 2



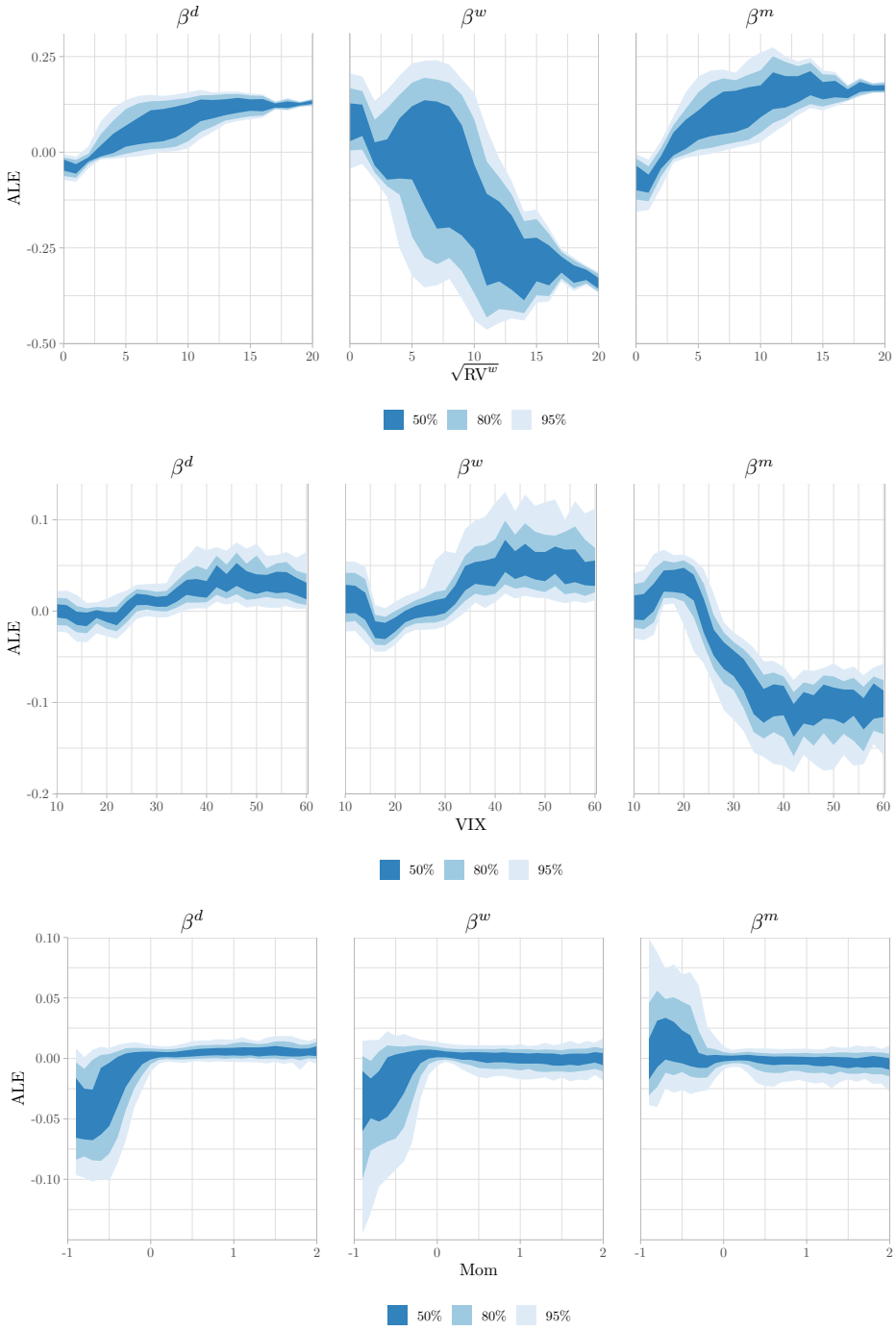
Notes: Visualizations of a series of splitting rules from Figure 2. Every split is aligned with one of the feature axes. Each split corresponds to drawing a line parallel to one of the axes. For a feature space of size 2, the space is divided into five regions, $(R_1, R_2, R_3, R_4, R_5)$, each of which is a two-dimensional rectangle. The parameters of the Panel-HAR model are estimated from the data of the corresponding rectangular. For example, all observations in the region R_5 will have the HAR coefficients $(\hat{\beta}_5^d, \hat{\beta}_5^w, \hat{\beta}_5^m)^T$. This concept generalizes straightforwardly to dimensions greater than two.

Figure 4: Time-varying HAR-Forest-Full coefficients



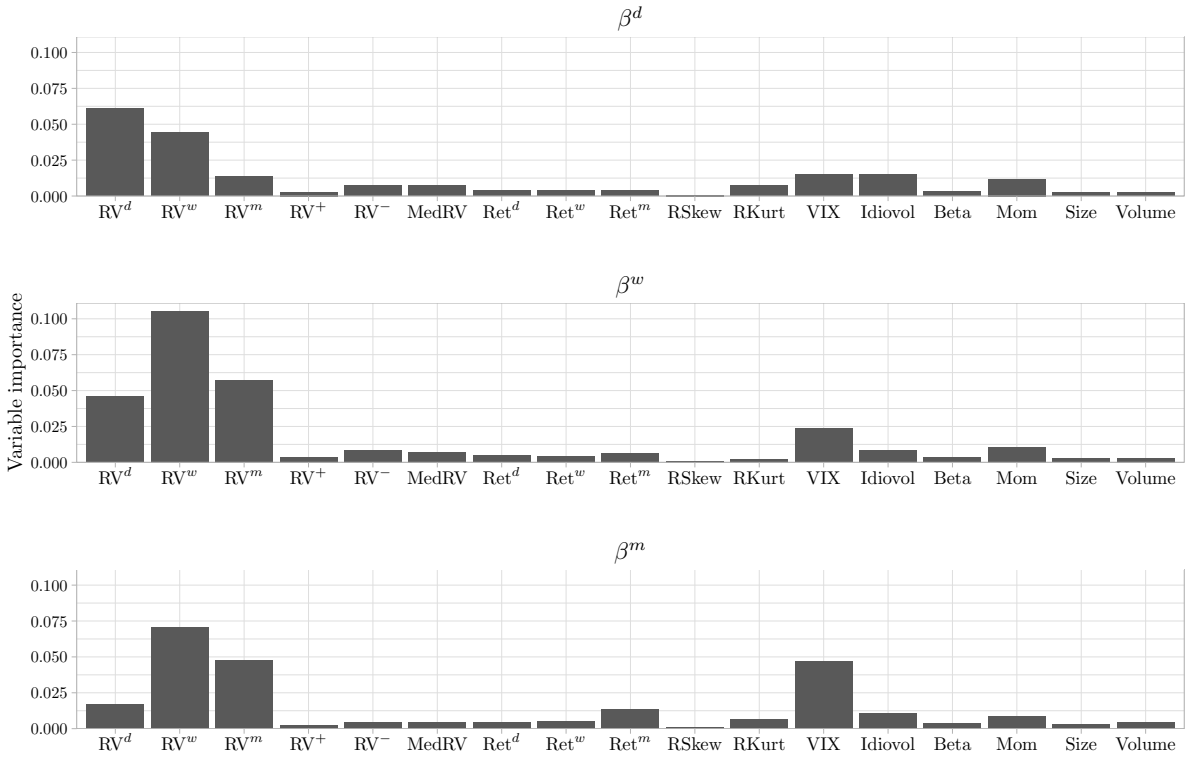
Notes: This figure depicts the bands containing 50%, 80%, and 95% of state-implied coefficients for the HAR-Forest-Full model estimated on the full sample targeting the 22-day-ahead forecast horizon. The fourth panel corresponds to the average of β^d, β^w, and β^m. Black horizontal lines depict the pooled full sample estimates from the Panel-HAR model.

Figure 5: Accumulated local effects per covariate



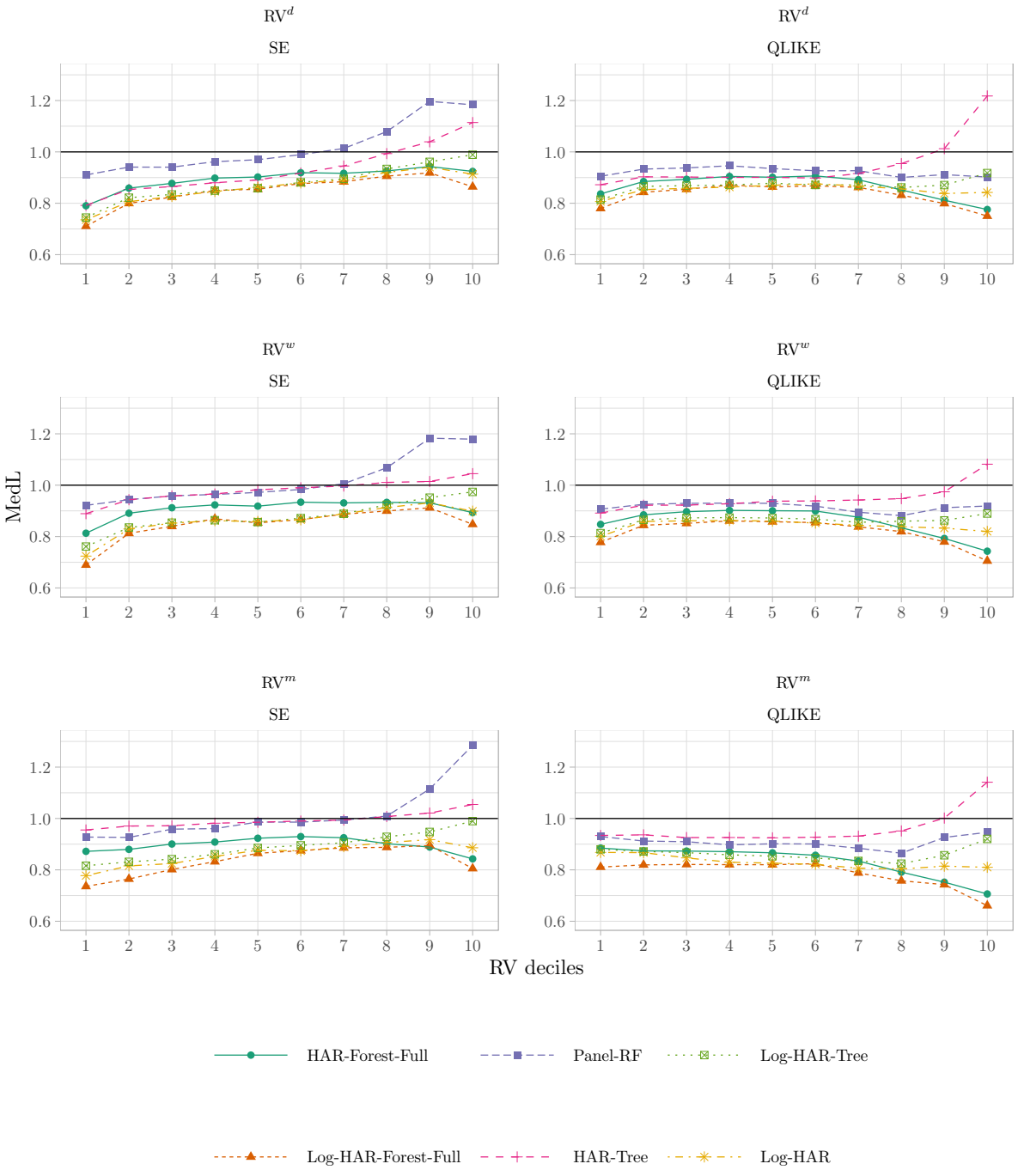
Notes: We calculate the ALE on HAR-Forest-Full parameters for all stocks and aggregate the individual curves to fan charts. The bands correspond to 50%, 80%, and 95% central intervals. The equally-spaced bins of the conditioned covariate in the upper row are $(, 1], \ldots, (19, 20]$, in the middle row they are $(10, 11], \ldots, (59, 60]$, and in the bottom row they are $(-1, -0.9], \ldots, (1.9, 2]$. Note that the ALE plots might not appear to be centered around zero because the x-axis is truncated.

Figure 6: Variable importance based on accumulated local effects



Notes: We calculate the variable importance per stock and depict the cross-sectional average variable importance for HAR-Forest-Full. The variable importance measures are calculated for all splitting variables except $\overline{RV}_i$ because the low number of different values for $\overline{RV}_i$ makes accumulated local effects unreliable. The histogram depicts the average ALE-implied variable importance across stocks.

Figure 7: Conditional median loss ratios per RV^d, RV^w, and RV^m deciles



Notes: We report the median loss ratio (MedL, see Eq. (8)) across different volatility regimes for the leading case of 22-day-ahead forecasts. For each stock and each right-hand side variable in the HAR model (rows 1–3), we group the OOS observations by deciles (1–10) of the corresponding right-hand-side variable in the HAR equation. MedL is calculated per decile across stocks. Note that this procedure ensures having the same number of observations per stock and decile.

Internet Appendix to

A Forest Full of Risk Forecasts for Managing Volatility

This Internet Appendix provides supplementary material to accompany the main paper. Section IA.1 defines the HARQ, SHAR, and HExp benchmark models. Section IA.2 defines the additional covariates used as splitting variables in the local linear forest model. Section IA.3 presents the Monte Carlo simulation study that evaluates the finite-sample performance of the HAR-Forest model under cross-sectional heterogeneity and macro-financial regimes. Sections IA.4 through IA.8 study the robustness of our findings to: (i) firm characteristics, (ii) longer forecast horizons, (iii) alternative lengths of the rolling estimation window, (iv) alternative HAR-Forest hyperparameter values, and (v) alternative machine-learning-inspired benchmark models.

IA.1 Definition of benchmark models: HARQ, SHAR, and HExp

Similar to our HAR model in Eq. (2), the HARQ model of Bollerslev et al. (2016) for stock i is defined as follows:

$$RV^{d}_{i,t+1} = \beta^{0,Q} + \beta^{d,Q} RV_{i,t} + \beta^{d,Q,inter} RV_{i,t} \times \sqrt{RQ_{i,t}} + \beta^{w,Q} RV^{w}_{i,t} + \beta^{m,Q} RV^{m}_{i,t} + \varepsilon^{Q}_{i,t}, \tag{IA1}$$

with $RQ_{i,t} = \frac{M}{3} \sum_{j=1}^{M} r^{4}_{t-(j-1)\cdot\Delta}$ and $\varepsilon^{Q}_{i,t}$ being an innovation term. In contrast to the standard HAR model, the HARQ includes an interaction term with the realized quarticity RQ.

The SHAR model of Patton and Sheppard (2015) extends the HAR framework by replacing the lagged daily RV by its two components; the realized semi-variance derived from positive returns and the realized semi-variance derived from negative returns:

$$RV^{d}_{i,t+1} = \beta^{0,S} + \beta^{d,S,+} RV^{+}_{i,t} + \beta^{d,S,-} RV^{-}_{i,t} + \beta^{w,S} RV^{w}_{i,t} + \beta^{m,S} RV^{m}_{i,t} + \varepsilon^{S}_{i,t}, \tag{IA2}$$

with $\varepsilon^{S}_{i,t}$ being an innovation term. For other forecast horizons beyond one day, we replace the left-hand sides of Eqs. (IA1) and (IA2) by the average RV of that forecast horizon. Estimation is carried out via weighted-least-squares regression as proposed in Patton and Sheppard (2015). We employ a two-stage estimation approach. First, we fit the model using ordinary-least-squares regression. In the second step, we apply weighted-least-squares estimation in which the weights are chosen to be inversely proportional to the first-stage predicted values.

The HExp model of Bollerslev et al. (2018) replaces the stepwise volatility components of the HAR model with a mixture of exponentially weighted moving averages of past realized volatilities. For a given

center of mass (CoM) and decay rate $\lambda = \log(1 + 1/\text{CoM})$, the exponential RV factor is

$$ExpRV_{i,t}^{\text{CoM}} \equiv \sum_{k=1}^{500} \frac{e^{-k\lambda}}{\sum_{s=1}^{500} e^{-s\lambda}} RV_{i,t+1-k}^{d} \,. \tag{IA3}$$

The HExp model employs four such factors with centers of mass equal to 1, 5, 25, and 125 trading days. Using the centering approach described in Section 2.1:

$$RV_{i,t+1:t+h} - \overline{RV_i} = \sum_{j \in \{1,5,25,125\}} \beta_j \big(ExpRV_{i,t}^{j} - \overline{RV_i} \big) + \varepsilon_{i,t} \,. \tag{IA4}$$

Because the EWMA factors are prespecified, the model contains no tuning parameters and is estimated by OLS. We estimate the HExp model on a stock-by-stock basis using the same rolling window scheme described in Section 3.2. Since the original model was developed for a panel of liquid futures contracts, applying it individually to a large cross-section of equities can occasionally produce extreme forecasts. Following the "insanity filter" logic in Bollerslev et al. (2018), we apply two in-sample-determined bounds: (i) a ratio cap constraining forecasts to $[\frac{1}{2}\,\overline{RV_i},\ 2\,\overline{RV_i}]$, and (ii) a clamp to the range of the target variable observed in the estimation window. Both bounds are fixed across all stocks and estimation periods and involve no out-of-sample tuning.

IA.2 Definition of additional covariates

We include additional splitting variables in our local linear forest based on daily stock-specific information derived from intraday data: semi-variances of positive and negative returns denoted by RV^{+} and RV^{-} (Barndorff-Nielsen et al., 2010; Patton and Sheppard, 2015), a jump-robust measure of volatility denoted by MedRV

$$RV_{i,t}^{+} = \sum_{j=1}^{M} r_{i,t-(j-1)\cdot\Delta}^{2} I_{\{r_{i,t-(j-1)\cdot\Delta} \geq 0\}} \text{ and } RV_{i,t}^{-} = \sum_{j=1}^{M} r_{i,t-(j-1)\cdot\Delta}^{2} I_{\{r_{i,t-(j-1)\cdot\Delta} < 0\}},$$

where $I_{\{\cdot\}}$ denotes the indicator function, and a jump-robust measure of volatility denoted by MedRV (Andersen et al., 2012),

$$MedRV_{i,t} = \frac{\pi}{6 - 4\sqrt{3} + \pi} \left(\frac{M}{M-2} \right) \sum_{j=2}^{M-1} \text{Median}(|r_{i,t-(j-1)\cdot\Delta}|, |r_{i,t-j\cdot\Delta}|, |r_{i,t-(j+1)\cdot\Delta}|).$$

Higher-order moments included in this analysis are the realized skewness and realized kurtosis (Amaya et al., 2015),

$$RSkew_{i,t} = \frac{\sqrt{M}\sum_{j=1}^{M} r_{i,t-(j-1)\cdot\Delta}^3}{RV_{i,t}^{3/2}} \text{ and } RKurt_{i,t} = \frac{M/3\sum_{j=1}^{M} r_{i,t-(j-1)\cdot\Delta}^4}{RV_{i,t}^2}.$$

IA.3 Simulation study

We construct a panel data-generating process (DGP) for realized volatility that is locally HAR but globally nonlinear, featuring cross-sectional heterogeneity and macro-financial regime dependence. The design provides a controlled environment in which the HAR-Forest's ability to recover state-dependent parameters can be directly assessed against the individual HAR and panel HAR benchmarks.

Sample size. The total simulation length is

$$T_{\text{total}} = T_{\text{obs}} + T_{\text{burnin}} + p_{\max},$$

where $T_{\text{burnin}} = 500$ eliminates dependence on initial conditions and $p_{\max} = 22$ corresponds to the monthly HAR component. The first 22 observations of each asset are initialized by drawing $RV_{j,t} \sim \mathcal{N}(10, 3^2)$, truncated below at zero to ensure non-negativity, and initial daily persistence coefficients drawn as $\beta_{j,t}^d \sim \mathcal{N}(0.3, 0.03^2)$, for $t = 1, \ldots, 22$.

Common volatility factor. The macro-financial state variable is the normalized VIX index. We use the actual historical VIX path, which is held fixed across all Monte Carlo replications. From the full history starting in 1990, we retain the last 2522 observations, corresponding to the period from June 23, 2016, to February 20, 2026. This segment serves as the true VIX path driving the data-generating process in the simulations. The total length $T_{total} = 2522$ reflects the simulation design and consists of 2000 effective observations, a burn-in period of 500 observations, and 22 lags used in model construction. We normalize it over the effective simulation window after burn-in:

$$z_t = \frac{VIX_t - \min_s VIX_s}{\max_s VIX_s - \min_s VIX_s}, \qquad z_t \in [0, 1], \tag{IA5}$$

where the min and max are taken over $s \in [T_{\text{burnin}}, T_{\text{total}}]$. Using the observed VIX path ensures that the simulated regimes mirror the macro-financial conditions present in our empirical application.

Panel structure and heterogeneity. The panel consists of N assets indexed by j. Each asset is assigned a fixed market exposure parameter β_j capturing differences in systematic risk. The exposure coefficients are constructed deterministically over the interval $[0.25, 1.75]$ by partitioning the interval into N equally spaced points. Specifically,

$$\beta_j = 0.25 + (j-1)\Delta, \qquad \Delta = \frac{1.75 - 0.25}{N-1}, \quad j = 1, \ldots, N, \tag{IA6}$$

so that assets span the full range from low- to high-market sensitivity. This design generates a smooth cross-sectional gradient in systematic exposure, ranging from defensive assets with low market beta to highly cyclical assets with strong comovement with aggregate market risk. Assets are partitioned into two groups: Group 1: $j \leq N/2$, Group 2: $j > N/2$, inducing systematic cross-sectional heterogeneity in volatility persistence.

Data-generating process. For $t > 22$, the HAR components are $RV^d_{j,t} = RV_{j,t-1}$, $RV^w_{j,t} = \frac{1}{5}\sum_{k=1}^{5} RV_{j,t-k}$, and $RV^m_{j,t} = \frac{1}{22}\sum_{k=1}^{22} RV_{j,t-k}$. Realized volatility evolves as

$$RV_{j,t} = c_{j,t} + \beta^d_{j,t} RV^d_{j,t} + \beta^w_{j,t} RV^w_{j,t} + \beta^m_{j,t} RV^m_{j,t} + \varepsilon_{j,t}, \qquad \varepsilon_{j,t} \sim \mathcal{N}(0, 0.03^2),$$

with innovations independent across j and t. The time-varying coefficients are specified as follows.

Intercept: $c_{j,t} \sim \mathcal{N}(1, 0.2^2)$ for Group 1 and $c_{j,t} \sim \mathcal{N}(4, 0.2^2)$ for Group 2.

Daily component: The daily coefficient depends nonlinearly on lagged volatility,

$$\beta^d_{j,t} = \tfrac{0.75}{20} RV_{j,t-1} + \xi^d_{j,t}, \qquad \xi^d_{j,t} \sim \mathcal{N}(0.7, 0.34^2), \quad j \leq N/2,$$
$$\beta^d_{j,t} = \tfrac{0.15}{50} RV_{j,t-1} + \xi^d_{j,t}, \qquad \xi^d_{j,t} \sim \mathcal{N}(0.2, 0.34^2), \quad j > N/2.$$

The scaling constants (20 and 50) are calibrated to keep the HAR recursion bounded while preserving state-dependent persistence.

Weekly component: The weekly coefficient is modulated by the VIX factor z_t:

$$\beta^w_{j,t} = \xi^w_{j,t}\, z_t, \quad \xi^w_{j,t} \sim \mathcal{N}(0.30, 0.14^2), \quad j \leq N/2,$$
$$\beta^w_{j,t} = \xi^w_{j,t}\, z_t, \quad \xi^w_{j,t} \sim \mathcal{N}(0.10, 0.14^2), \quad j > N/2,$$

so that medium-horizon persistence strengthens during periods of elevated market uncertainty.

Monthly component: $\beta^m_{j,t} \sim \mathcal{N}(0.10, 0.07^2)$ for Group 1 and $\beta^m_{j,t} \sim \mathcal{N}(0.30, 0.07^2)$ for Group 2.

The resulting panel exhibits nonlinear, state-dependent HAR persistence, macro-financial regime effects driven by the VIX factor, and systematic cross-sectional heterogeneity.

Estimation. We compare three models: the individual HAR, estimated separately for each asset; the panel HAR, estimated by pooling all assets; and the HAR-forest, which employs lagged daily RV, z_t, a proxy for market beta (eq. IA6), and weekly and monthly lagged volatility as candidate splitting variables in order to recover the regime- and cross-sectional-heterogeneity-dependent structure of the DGP. Both HAR benchmarks are estimated as described in in Section 2, with asset-level demeaning applied throughout. Notably, the weekly and monthly lagged volatility measures are used solely as potential partitioning variables and do not directly affect the simulation DGP.

Results. Table IA.1 reports five-number summaries of the MSE distributions across 500 independent Monte Carlo replications for $N \in \{10, 20, 40\}$. The HAR-Forest consistently achieves lower MSE than both benchmarks across all panel sizes, with MSE ratios relative to the individual HAR ranging from 0.97 to 0.99 for $N = 10$ and improving to 0.95–0.97 for $N = 40$, indicating that the advantage grows with the size of the cross-section. The individual HAR and panel HAR deliver near-identical MSE throughout, as expected given that the DGP features time-varying rather than stock-fixed heterogeneity, which limits the gains from pooling alone. Since each replication constitutes an independent draw, standard two-sample t-tests comparing the HAR-Forest against the panel HAR are appropriate; the resulting p-values are 0.0006, $< 10^{-16}$, and 0.035 for $N = 10$, $N = 20$, and $N = 40$, respectively, confirming that the MSE

differences are statistically significant across panel sizes.

Table IA.1: Forecast errors across simulation replications

Statistic	Individual HAR	Panel HAR	HAR forest	Ratio (HAR forest / Individual HAR)
Panel size $N = 10$				
Min	7.22	7.22	7.07	0.97
Q1	7.89	7.88	7.77	0.98
Median	8.11	8.10	7.98	0.99
Q3	8.34	8.33	8.23	0.99
Max	9.29	9.29	9.21	1.00
Panel size $N = 20$				
Min	7.50	7.50	7.22	0.95
Q1	7.99	7.98	7.70	0.96
Median	8.15	8.14	7.87	0.96
Q3	8.34	8.33	8.04	0.97
Max	8.87	8.87	8.58	0.98
Panel size $N = 40$				
Min	7.70	7.69	7.36	0.95
Q1	8.06	8.05	7.74	0.96
Median	8.18	8.16	7.86	0.96
Q3	8.29	8.28	7.96	0.97
Max	8.73	8.72	8.40	0.97

Notes: Five-number summaries of MSE across 500 Monte Carlo replications. The last column reports the ratio of HAR-Forest MSE to individual HAR MSE.

IA.4 Relation to stock characteristics

We explore whether our improvements in forecast performance can be related to certain firm characteristics. For that purpose, we move away from joint forecast evaluation via MCSs and compare individual HAR forecasts and our HAR-Forest-Full using the Diebold-Mariano (DM) test (Diebold and Mariano, 1995). For this analysis, we merge the DM statistics with additional firm characteristics from Gu et al. (2020): book-to-market ratios, capital expenditures and inventory, cash holdings, earnings announcement returns, financial statement score, illiquidity, leverage, share turnover, size, and trading volume.[20] As the DM test statistics are based on the out-of-sample period 2005–2021, firm characteristics are averaged over the same period.

Regression results of the DM test statistics on the characteristics are reported in Table IA.2. The intercept is always positive and statistically significant with t-statistics larger than 5, indicating that the

[20]The selection of firm characteristics is by no means extensive. We could include more characteristics from Gu et al. (2020), but this would make the interpretation less transparent. Hence, we restrict ourselves to a (subjective) selection of characteristics.

HAR-Forest-Full outperforms the individually estimated HAR models across stocks. If we add the firm characteristics one-by-one, only cash holdings have a statistically significant explanatory effect on loss differences at the 1% significance level. However, even in this case the R^2 is only equal to 1.34%.

If we now consider a joint regression of all firm characteristics on the forecast evaluation outcome, the adjusted R^2 is higher than in the regression that includes only cash holdings but still lower than 2.1%. As a robustness check, we split the full sample into two equal time intervals and run linear regressions for both samples. The first sample is from 2005 to 2013M6 and includes the global financial crisis, whereas the second sample from 2013M7 to 2021 includes the covid-19 crisis period. The intercept term remains statistically significant in both subsamples with t-statistics larger than 5.9 and the R^2s of these regressions remain small as the maximum R^2 is only 2.49%. Hence, we conclude that our HAR-Forest-Full performs well across firm characteristics.

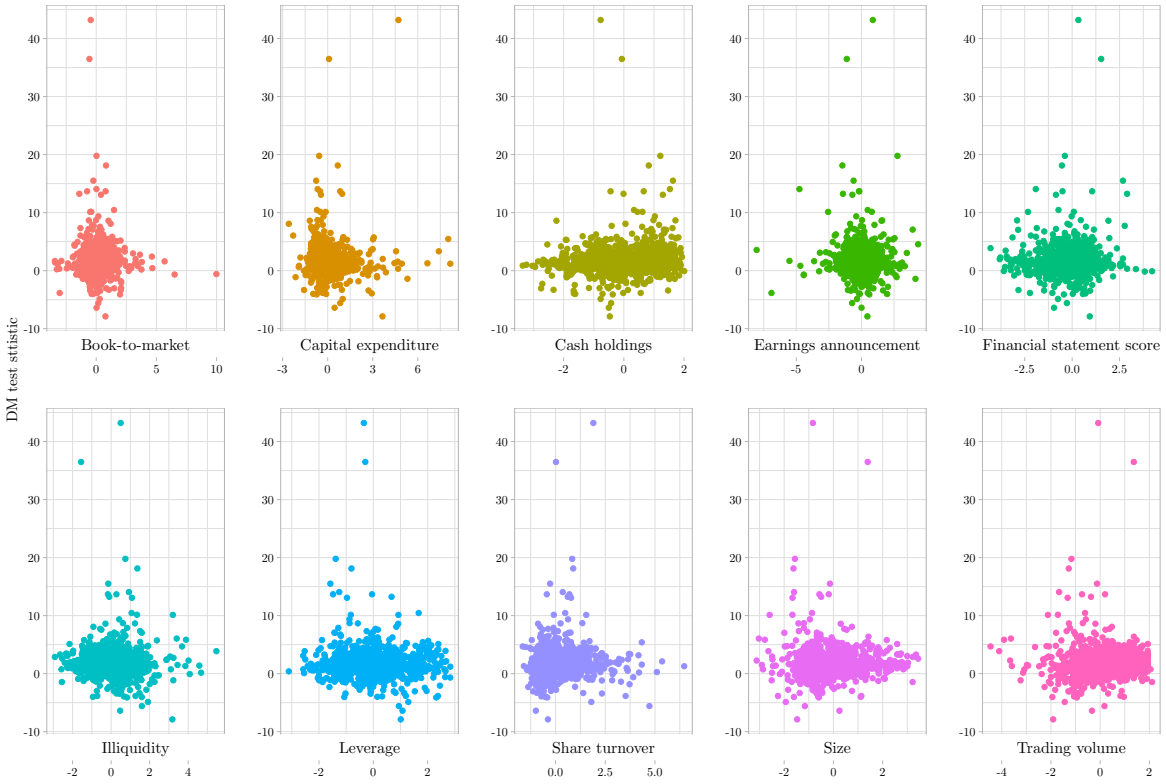
Figure IA.1 illustrates how the DM test statistics vary with each of the firm characteristics. Overall, the scatter plots confirm that the majority of observations lie above zero, reflecting the HAR-Forest-Full's consistent outperformance relative to the individually estimated HAR models. At the same time, most characteristics do not exhibit a pronounced linear relationship with the DM statistics, as evidenced by the largely diffuse scatter. Consequently, the figure supports our conclusion that the HAR-Forest-Full delivers improvements in predictive accuracy that hold even after accounting for key firm-level characteristics.

Table IA.2: Forecast comparison and firm characteristics

	Full sample												2005M1–2013M6	2013M7–2021M12
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(Intercept)	1.77***	1.77***	1.78***	1.77***	1.77***	1.77***	1.77***	1.77***	1.77***	1.77***	1.77***	1.78***	2.11***	1.48***
	(22.09)	(22.08)	(21.93)	(22.23)	(22.06)	(22.07)	(22.08)	(22.09)	(22.10)	(22.06)	(22.08)	(22.00)	(17.32)	(5.97)
Book-to-market	—	-0.05	—	—	—	—	—	—	—	—	—	0.06	-0.14	-0.16
		(-0.74)										(0.65)	(-0.94)	(-1.45)
Capital expenditure	—	—	-0.09	—	—	—	—	—	—	—	—	0.01	-0.08	-0.29
			(-0.45)									(0.06)	(-0.30)	(-1.06)
Cash holdings	—	—	—	0.32***	—	—	—	—	—	—	—	0.45***	0.65***	-0.14
				(4.39)								(4.60)	(5.26)	(-0.83)
Earnings announcement	—	—	—	—	-0.09	—	—	—	—	—	—	-0.14	0.04	0.21
					(-0.75)							(-1.12)	(0.29)	(0.81)
Financial statement score	—	—	—	—	—	0.06	—	—	—	—	—	0.17	0.27*	-0.28
						(0.63)						(1.45)	(1.90)	(-1.34)
Illiquidity	—	—	—	—	—	—	-0.11	—	—	—	—	-0.60**	0.03	-0.94
							(-1.18)					(-2.18)	(0.08)	(-1.09)
Leverage	—	—	—	—	—	—	—	-0.12	—	—	—	0.08	0.26	-0.01
								(-1.57)				(0.81)	(1.57)	(-0.05)
Share turnover	—	—	—	—	—	—	—	—	0.17*	—	—	0.10	0.17	0.45
									(1.70)			(0.57)	(0.55)	(0.81)
Size	—	—	—	—	—	—	—	—	—	-0.00	—	-0.27	-0.05	-1.05*
										(-0.01)		(-0.79)	(-0.11)	(-1.65)
Trading volume	—	—	—	—	—	—	—	—	—	—	0.08	-0.25	-0.03	0.20
											(0.96)	(-0.71)	(-0.05)	(0.44)
Adj. R^2	—	-0.0006	0.0002	0.0134	0.0001	-0.0004	0.0007	0.0010	0.0031	-0.0009	0.0000	0.0209	0.0249	-0.0002

Notes: We regress the Diebold–Mariano test statistics of equal predictive ability on average firm characteristics computed over the OOS period 2005–2021. The tests compare forecasts of individually-estimated HAR models vs. the HAR-Forest (100). A positive test statistic indicates that the individually-estimated HAR model predictions are outperformed by the predictions derived from the HAR forests. The firm characteristics are standardized to have mean zero and unit variance. For the definition of the firm characteristics see Gu et al. (2020). Cash holdings, illiquidity, and leverage are logarithmized before standardization to match Figure IA.1. The test statistics are based on heteroscedasticity-robust standard errors. In the last two columns, we report the same analysis as in regression (12) but on two disjoint subsamples of the OOS period.

Figure IA.1: Diebold-Mariano test statistics relative to standardized firm characteristics



Notes: We depict scatter plots of average firm characteristics in the out-of-sample period on the Diebold-Mariano test statistic when comparing forecasts based on individual HAR models with the average forest forecast for the 22-day-ahead horizon. Positive test statistics indicate superior forecast performance of the HAR-Forest-Full. The firm characteristics are standardized to have mean zero and unit standard deviation. The characteristics are taken from Gu et al. (2020) but cash holdings, illiquidity, and leverage are logarithmized before standardization. The OOS forecast period is 2005–2021.

IA.5 Longer forecast horizons

Table IA.3 presents the forecast evaluation results for longer forecast horizons (44-day-ahead and 66-day-ahead forecasts in Panels A and B, respectively). Across both panels, the HAR-Forest-based models consistently deliver lower MedL for both the SE and QLIKE criteria compared to the alternative models. In the case of 44-day-ahead forecasts, the Log-HAR-Forest-Full model achieves the lowest SE-MedL ratio (0.886), while the HAR-Forest-RV and Log-HAR-Forest-RV models also perform strongly across SE, QLIKE, and utility metrics. These models also maintain higher inclusion rates with respect to competing specifications. The performance advantages of the HAR-Forest-based models also persist for 66-day-ahead forecasts. The Log-HAR-Forest-Full and Log-HAR-Forest-RV models, in particular, continue to exhibit superior forecasting accuracy by achieving the best or near-best performance in both the SE and QLIKE metrics, while also delivering competitive utility gains.

Table IA.3: Robustness check: Longer forecasting horizons

	SE		QLIKE		Utility	
Model	MedL	MCSR	MedL	MCSR	MedU	MCSU
Panel D: 44-day-ahead						
HAR-Forest-Full	0.917	0.680	0.842	0.638	1.036	0.716
HAR-Forest-RV	0.974	0.592	0.813	0.710	**1.036**	0.759
Log-HAR-Forest-Full	**0.886**	**0.858**	0.818	0.828	1.027	**0.839**
Log-HAR-Forest-RV	0.902	0.812	**0.804**	**0.829**	1.030	0.830
Panel-RF	1.166	0.460	0.927	0.486	1.016	0.617
HAR-Tree	1.000	0.504	0.976	0.368	1.000	0.502
Log-HAR-Tree	0.940	0.754	0.880	0.664	1.013	0.725
Panel-HAR	0.997	0.574	1.017	0.329	0.999	0.440
HARQ	0.986	0.528	0.987	0.353	1.001	0.462
SHAR	0.953	0.654	0.914	0.513	1.013	0.625
Log-HAR	0.926	0.800	0.879	0.702	1.013	0.762
HExp	0.991	0.671	0.974	0.493	1.001	0.591
Equal-Avg	0.901	0.713	0.854	0.613	1.025	0.719
HAR	—	0.530	—	0.336	—	0.459
Panel E: 66-day-ahead						
HAR-Forest-Full	0.979	0.540	0.877	0.600	1.024	0.665
HAR-Forest-RV	1.037	0.486	0.847	0.687	**1.030**	0.726
Log-HAR-Forest-Full	0.914	**0.808**	0.852	**0.817**	1.022	**0.814**
Log-HAR-Forest-RV	0.941	0.718	**0.830**	0.803	1.025	0.799
Panel-RF	1.097	0.433	0.937	0.529	1.014	0.635
HAR-Tree	1.004	0.432	0.973	0.399	1.002	0.489
Log-HAR-Tree	0.946	0.694	0.888	0.693	1.012	0.718
Panel-HAR	0.978	0.579	1.005	0.405	1.000	0.479
HARQ	0.988	0.464	0.984	0.362	1.001	0.466
SHAR	0.953	0.607	0.920	0.541	1.012	0.621
Log-HAR	**0.907**	0.768	0.890	0.720	1.012	0.750
HExp	0.955	0.683	0.958	0.552	1.004	0.623
Equal-Avg	0.910	0.667	0.870	0.633	1.022	0.694
HAR	—	0.505	—	0.353	—	0.445

Notes: We report the median loss ratio (MedL) (lower is better), and the MCS inclusion rate (MCSR) across stock (higher is better); see Section 3.5. The last three columns report similar figures for the utility framework by Bollerslev et al. (2018). We report the median utility ratio (MedU) relative to the individual HAR benchmark model (higher is better), and the MCS inclusion rate of the utility measure (MCSU) across stocks (higher is better). Results for individually fitted HAR models are found in the rows labelled "HAR." The number of trees per HAR-Forest is 200. The number of trees in the random forest is 500. The forecasts are issued daily and, hence, overlapping. The OOS period is 2005–2021. Averages are taken across the 500 stocks in each estimation window. Numbers in bold indicate the best outcome per panel and evaluation criterion (lowest for MedL, highest for MCSR, MedU, and MCSU). The MCS inclusion rates are based on the models listed in the table and all other benchmark models listed in Section IA.8.

IA.6 Length of the rolling estimation window

Chassot and Audrino (2025) find that the choice of the length of the rolling estimation window can have profound effects on the performance of HAR models, with larger training windows generally leading to lower prediction errors. Inspired by these findings, we conduct a robustness analysis by comparing two estimation window sizes for 22-day-ahead forecasts while keeping the out-of-sample period fixed at

2005–2021. Specifically, we examine a shorter three-year window and a longer ten-year window. For the latter, we extend our dataset back to 1995 to ensure a full ten-year estimation window at the start of the out-of-sample period. As before, we use a rolling window with yearly reestimation.

Since filtering for stocks with complete observations over 3-year or 10-year estimation windows alters the sample composition compared to the 5-year window used in the main analysis, we limit our focus to a subset of models: the (Log-)HAR-Forest models with either all covariates or only RV as splitting variables, along with the Log-HAR and HAR models as individually estimated benchmarks.

We report the forecast evaluation outcome for these alternative estimation windows in Table IA.4. Across both estimation windows, the HAR-Forest-Full and Log-HAR-Forest-Full models continue to outperform alternative specifications, reinforcing their robustness to variations in estimation window length. This is consistent with the results in Table 3, where these models exhibited the best MedL and MSCR for SE and QLIKE loss. For some HAR-Forest models, we even observe a decrease in MedL when going from a ten-year to a three-year estimation window. This may happen because we also reestimate the benchmark with only three years' worth of data when going from Panel A to Panel B. Because the HAR-Forest exploits cross-sectional information, we interpret this as evidence that HAR forests are more robust to short estimation windows than individually estimated HAR models.

Table IA.4: Robustness check: Estimation windows with different lengths

	SE		QLIKE		Utility	
Model	MedL	MCSR	MedL	MCSR	MedU	MCSU
Panel A: Ten-year estimation window						
HAR-Forest-Full	0.895	0.809	0.825	0.793	**1.023**	0.875
HAR-Forest-RV	0.956	0.679	0.847	0.698	**1.023**	0.827
Log-HAR-Forest-Full	**0.826**	**0.988**	**0.802**	**0.959**	1.019	**0.949**
Log-HAR-Forest-RV	0.846	0.881	0.812	0.850	1.019	0.905
Log-HAR	0.863	0.856	0.844	0.762	1.016	0.840
HAR	—	0.506	—	0.228	—	0.438
Panel B: Three-year estimation window						
HAR-Forest-Full	0.886	0.807	0.832	0.678	1.030	0.717
HAR-Forest-RV	0.912	0.748	0.810	0.764	**1.035**	0.817
Log-HAR-Forest-Full	**0.881**	**0.922**	0.798	0.828	1.027	0.785
Log-HAR-Forest-RV	0.885	0.893	**0.794**	**0.864**	1.029	**0.835**
Log-HAR	0.947	0.822	0.882	0.589	1.014	0.635
HAR	—	0.597	—	0.340	—	0.451

Notes: We report the median loss ratio (MedL) (lower is better) and the MCS inclusion rate (MCSR) across stock (higher is better); see Section 3.5. The last two columns report similar figures for the utility framework by Bollerslev et al. (2018). We report the median utility ratio (MedU) relative to the individual HAR benchmark model (higher is better), and the MCS inclusion rate of the utility measure (MCSU) across stocks (higher is better). Results for individually fitted HAR models are found in the rows labelled "HAR." The OOS period is 2005–2021. Numbers in bold indicate the best outcome per panel and evaluation criterion (lowest for MedL, highest for MCSR, MedU, and MCSU).

IA.7 Alternative HAR-Forest hyperparameter values

In the following, we explore the sensitivity of the HAR-Forest model's performance to key hyperparameter choices. First, we examine the effect of reducing the number of HAR-Forest trees from 200 to 50. Second, we investigate the impact of increasing the minimum node size from 100 to 500. Finally, we assess the consequences of changing the feature subsampling rate from 1/3 to 1/2.

IA.7.1 Number of HAR-Forest trees

Table IA.5 presents the forecast evaluation results when the HAR-Forest model is configured with 50 trees instead of the standard 200. Across all forecast horizons, the Log-HAR-Forest-Full model consistently achieves the best or near-best performance in terms of SE-MedL and QLIKE-MedL, while the HAR-Forest-Full model remains competitive and often leads in utility metrics despite the reduction in the number of trees.

Table IA.5: Robustness check: HAR-Forest model with 50 trees instead of 200 trees

	SE		QLIKE		Utility	
Model	MedL	MCSR	MedL	MCSR	MedU	MCSU
Panel A: 1-day-ahead						
HAR-Forest-Full	0.947	0.603	0.858	0.336	**1.029**	**0.400**
HAR-Forest-RV	0.959	0.521	0.900	0.193	1.018	0.339
Log-HAR-Forest-Full	**0.925**	**0.869**	**0.844**	**0.674**	1.015	0.365
Log-HAR-Forest-RV	0.932	0.788	0.854	0.473	1.012	0.305
Panel B: 5-day-ahead						
HAR-Forest-Full	0.860	0.714	0.827	0.659	**1.028**	0.804
HAR-Forest-RV	0.894	0.614	0.858	0.463	1.023	0.669
Log-HAR-Forest-Full	**0.847**	**0.935**	**0.799**	**0.907**	1.022	**0.821**
Log-HAR-Forest-RV	0.865	0.809	0.808	0.730	1.021	0.728
Panel C: 22-day-ahead						
HAR-Forest-Full	0.904	0.711	0.832	0.682	1.026	0.771
HAR-Forest-RV	0.945	0.635	0.818	0.698	**1.029**	0.770
Log-HAR-Forest-Full	**0.868**	**0.905**	0.814	0.852	1.023	0.857
Log-HAR-Forest-RV	0.887	0.841	**0.806**	**0.871**	1.024	**0.873**

Notes: See notes to Table IA.4, except that we use 50 trees instead of 200 trees compared to the main analysis with a five-year estimation window.

IA.7.2 Minimum node size

The evaluation presented in Table IA.6 demonstrates that increasing the minimum node size from 100 to 500 does not adversely affect the model's performance. Across forecast horizons, the Log-HAR-Forest-Full

model consistently delivers lower median loss ratios for both SE and QLIKE metrics while achieving the highest MCS inclusion rates. This indicates that a more conservative node splitting criterion—resulting from a larger minimum node size—does not come at the cost of sacrificing forecast accuracy.

Table IA.6: Robustness check: HAR-Forest model with minimum node size of 500 instead of 100

	SE		QLIKE		Utility	
Model	MedL	MCSR	MedL	MCSR	MedU	MCSU
Panel A: 1-day-ahead						
HAR-Forest-Full	0.962	0.506	0.876	0.175	1.025	0.233
HAR-Forest-RV	0.966	0.501	0.878	0.185	**1.029**	**0.363**
Log-HAR-Forest-Full	**0.932**	**0.840**	**0.847**	**0.620**	1.015	0.326
Log-HAR-Forest-RV	0.935	0.788	0.854	0.477	1.013	0.280
Panel B: 5-day-ahead						
HAR-Forest-Full	0.895	0.646	0.848	0.452	1.023	0.670
HAR-Forest-RV	0.911	0.569	0.852	0.461	**1.025**	0.716
Log-HAR-Forest-Full	**0.838**	**0.939**	**0.797**	**0.908**	1.023	**0.842**
Log-HAR-Forest-RV	0.851	0.857	0.800	0.791	1.022	0.809
Panel C: 22-day-ahead						
HAR-Forest-Full	0.914	0.684	0.851	0.588	1.023	0.704
HAR-Forest-RV	0.948	0.597	0.822	0.676	**1.031**	0.759
Log-HAR-Forest-Full	**0.855**	**0.914**	0.805	0.856	1.026	0.863
Log-HAR-Forest-RV	0.868	0.832	**0.797**	**0.871**	1.027	**0.882**

Notes: See notes to Table IA.4, expect that we use a minimum node size of 500 instead of 100 compared to the main analysis with a five-year estimation window.

IA.7.3 Feature subsampling rate

Adjusting the feature subsampling rate from the default 1/3 to 1/2 yields robust forecasting performance across various horizons, as per Table IA.7. The Log-HAR-Forest-Full model maintains strong results, with lower median loss ratios and high inclusion rates under both SE and QLIKE evaluations, and the HAR-Forest-Full model delivers competitive utility measures. This finding suggests that employing a higher subsampling rate, which incorporates a larger set of features at each split, does not lead to a deterioration in forecast performance.

Table IA.7: Robustness check: HAR-Forest model with feature subsampling rate equal to 1/2 instead of 1/3

	SE		QLIKE		Utility	
Model	MedL	MCSR	MedL	MCSR	MedU	MCSU
Panel A: 1-day-ahead						
HAR-Forest-Full	0.942	0.630	0.850	0.365	**1.031**	**0.447**
HAR-Forest-RV	0.953	0.571	0.878	0.251	1.028	0.414
Log-HAR-Forest-Full	**0.924**	**0.873**	**0.846**	**0.668**	1.015	0.371
Log-HAR-Forest-RV	0.932	0.768	0.854	0.478	1.012	0.296
Panel B: 5-day-ahead						
HAR-Forest-Full	0.858	0.728	0.820	0.673	**1.028**	**0.813**
HAR-Forest-RV	0.885	0.641	0.843	0.522	1.027	0.744
Log-HAR-Forest-Full	**0.846**	**0.939**	**0.798**	**0.896**	1.022	0.805
Log-HAR-Forest-RV	0.865	0.800	0.808	0.740	1.021	0.725
Panel C: 22-day-ahead						
HAR-Forest-Full	0.905	0.706	0.832	0.663	1.027	0.760
HAR-Forest-RV	0.941	0.633	0.819	0.697	**1.030**	0.765
Log-HAR-Forest-Full	**0.872**	**0.911**	0.814	0.831	1.023	0.840
Log-HAR-Forest-RV	0.887	0.841	**0.806**	**0.854**	1.024	**0.858**

Notes: See notes to Table IA.4, expect that we use a subsampling rate of 1/2 instead of 1/3 compared to the main analysis with a five-year estimation window.

IA.8 Alternative machine-learning inspired models

We benchmark the performance of our HAR-Forest model against two alternative classes of ML-inspired models. First, we consider models that incorporate cross-sectional volatility spillovers by allowing the realized variance of one stock to be influenced by the lagged variances of all other stocks. Second, we consider models that relax the linearity assumption present in the traditional HAR framework by adopting more flexible nonlinear machine learning models such as gradient boosting and neural networks.[21] All additional benchmark models are individually estimated for each stock.

IA.8.1 Model descriptions

Models that incorporate volatility spillover effects

AR: The linear AR model is given by:

$$RV_{t+1}^{d,i} = \alpha^{(i)} + \sum_{j=1}^{N} \beta_j^{(i)} RV_t^{d,j} + \varepsilon_t^{(i)}, \quad i = 1, \ldots, N \tag{IA7}$$

[21] Neural networks for volatility forecasting are also employed by Carr et al. (2020); Bucci (2020); Rahimikia and Poon (2020), among others.

where $RV_t^{d,j}$ represents the realized variance of stock j at time t, $\alpha^{(i)}$ is an intercept term, $\beta_j^{(i)}$ denotes the spillover effect from stock j to stock i, and $\varepsilon_t^{(i)}$ is an innovation term.

AR-Lasso: The lasso-regularized version of Eq. (IA7):

$$RV_{t+1}^{d,i} = \alpha^{(i)} + \sum_{j=1}^{N} \beta_j^{(i)} RV_t^{d,j} + \varepsilon_t^{(i)}, \quad \text{subject to} \quad \sum_{j=1}^{N} |\beta_j^{(i)}| \leq \lambda \tag{IA8}$$

where λ is a tuning parameter that controls the amount of regularization, effectively shrinking some of the coefficients $\beta_j^{(i)}$ to zero and thus selecting only the most important spillover effects.

AR-XGBoost: We consider a non-linear gradient boosting model, which allows for complex interactions between the realized volatilities of different stocks:

$$RV_{t+1}^{d,i} = f\left(\{RV_t^{d,j}\}_{j=1}^{N}\right) + \varepsilon_t^{(i)}, \tag{IA9}$$

where $f(\cdot)$ represents a non-linear function learned through gradient boosting.

Models that relax the linearity assumption in the HAR framework

Gradient boosting-HAR: In this model, we employ a gradient boosting algorithm to model the realized variance using the heterogeneous lagged components defined by the HAR framework. The features include lagged realized variances on daily, weekly, and monthly horizons (RV_t^d, RV_t^w, RV_t^m), as in the traditional HAR model, but instead of relying on a linear relationship, we use gradient boosting to flexibly model these lagged components.

Random forest-HAR: This variant applies a random forest to the HAR regressors. The regressors are the same as those used in the gradient boosting-HAR, which include daily, weekly, and monthly realized variance components.

Neural network-HAR: The neural network-HAR model uses a feedforward neural network to capture non-linear dependencies in the HAR lagged structure. We employ a feedforward neural network with ReLU activation functions and at most three hidden layers. The network is trained using the Adam optimizer with an initial learning rate of 0.001 and early stopping based on a 20% validation split of the training window. We perform a grid search over architectures with 1–3 hidden layers and 16, 32, or 64 units per layer, selecting the configuration that minimizes validation loss. The inputs are the same lagged realized volatility components used in the other HAR variants: daily, weekly, and monthly realized variances.

IA.8.2 Forecast evaluation results

Table IA.8 presents the forecast evaluation results for the additional benchmark models. Overall, the Log-HAR-XGBoost method achieves the best performance across most forecast horizons and evaluation criteria, with few exceptions. However, its performance falls short of the standard HAR model in terms of SE-MedL, QLIKE-MedL, and utility-MedU. Additionally, the inclusion rates of Log-HAR-XGBoost are often comparable to, or even lower than, those achieved by the traditional HAR model. These results confirm that our proposed HAR-Forest models also outperform a broader class of ML-inspired models. This finding is somewhat at odds with prior studies documenting strong ML performance. We attribute this discrepancy to two factors. First, we have a different sample both in terms of time dimension (including the COVID-19 crisis) and in terms of a larger cross-section that includes smaller stocks than those in the Dow Jones Industrial Average. Second, we include the overnight to capture the full-day variation of stock returns. However, this makes the realized variance proxy somewhat noisier as the entire overnight variation is encoded in only one squared overnight return. This may cause more flexible HAR-type machine learning models to overfit the noise.

Table IA.8: Robustness check: Additional benchmark models

	SE		QLIKE		Utility	
Model	MedL	MCSR	MedL	MCSR	MedU	MCSU
Panel A: 1-day-ahead						
AR	4.291	0.036	12.145	0.001	-3.755	0.006
Log-AR	1.070	0.151	1.533	0.025	0.814	0.042
AR-Lasso	1.205	0.047	2.036	0.007	0.613	0.020
Log-AR-Lasso	1.094	0.119	1.616	0.050	0.715	0.053
AR-XGBoost	1.201	0.087	1.773	0.025	0.675	0.036
Log-AR-XGBoost	1.035	0.219	1.245	0.071	0.866	0.064
HAR-XGBoost	1.158	0.099	1.798	0.027	0.632	0.025
Log-HAR-XGBoost	1.038	0.213	1.162	0.069	0.878	0.057
HAR-RF	1.418	0.038	2.168	0.007	0.602	0.024
Log-HAR-RF	1.242	0.043	1.999	0.012	0.610	0.030
HAR-NN	1.464	0.023	2.633	0.006	0.438	0.017
Log-HAR-NN	1.254	0.051	2.084	0.014	0.578	0.027
Panel B: 5-day-ahead						
AR	3.486	0.075	7.398	0.017	-0.850	0.035
Log-AR	1.126	0.177	1.525	0.064	0.865	0.097
AR-Lasso	1.355	0.098	2.727	0.050	0.526	0.080
Log-AR-Lasso	1.183	0.201	1.624	0.131	0.803	0.160
AR-XGBoost	1.225	0.180	1.290	0.115	0.928	0.157
Log-AR-XGBoost	1.077	0.322	1.231	0.165	0.913	0.163
HAR-XGBoost	1.231	0.164	1.342	0.097	0.891	0.139
Log-HAR-XGBoost	1.096	0.279	1.180	0.143	0.922	0.165
HAR-RF	1.660	0.056	2.204	0.043	0.732	0.085
Log-HAR-RF	1.497	0.075	2.105	0.058	0.730	0.089
HAR-NN	1.613	0.066	2.556	0.047	0.630	0.081
Log-HAR-NN	1.532	0.093	2.145	0.064	0.716	0.100
Panel C: 22-day-ahead						
AR	2.414	0.115	5.573	0.038	-0.275	0.081
Log-AR	1.230	0.199	1.496	0.117	0.893	0.182
AR-Lasso	1.299	0.192	2.121	0.111	0.718	0.166
Log-AR-Lasso	1.161	0.371	1.444	0.213	0.896	0.296
AR-XGBoost	1.242	0.231	1.238	0.195	0.947	0.267
Log-AR-XGBoost	1.128	0.410	1.264	0.246	0.923	0.287
HAR-XGBoost	1.195	0.285	1.203	0.229	0.949	0.315
Log-HAR-XGBoost	1.101	0.430	1.149	0.266	0.957	0.340
HAR-RF	1.678	0.146	1.890	0.106	0.843	0.198
Log-HAR-RF	1.489	0.191	1.789	0.129	0.839	0.206
HAR-NN	1.627	0.170	2.116	0.118	0.776	0.189
Log-HAR-NN	1.499	0.223	1.833	0.147	0.828	0.213

Notes: We report the median loss ratio (MedL) relative to the HAR benchmark model (lower is better) and the MCS inclusion rate (MCSR) across stocks (higher is better); see Section 3.5 and Table 3. The last two columns report similar figures for the utility framework by Bollerslev et al. (2018). We report the median utility ratio (MedU) relative to the individual HAR benchmark model (higher is better), and the MCS inclusion rate of the utility measure (MCSU) across stocks (higher is better). The OOS period is 2005–2021.