Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

Linear and nonlinear econometric models against machine learning models: realized volatility prediction

Rehim Kilic

2025-061

Please cite this paper as:
Kilic. Rehim (2026). "Linear and nonlinear econometric models against machine learning models: realized volatility prediction," Finance and Economics Discussion Series 2025-061r1. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2025.061r1.

Linear and Nonlinear Econometric Models versus Machine-Learning Models: Evidence from Realized-Volatility Forecasting

Rehim Kılıç*

August 2026

Abstract

This paper examines which representations of persistence and nonlinearity are most useful for forecasting realized volatility and whether machine learning adds value beyond econometric models designed for long memory and regime dependence. We compare HAR, ARFIMA, threshold HAR, smooth-transition HAR, and Markov-switching HAR with XGBoost and several neural-network models for the S&P 500 and 40 U.S. equities. Models are evaluated under a baseline information set using realized-volatility history and an extended set of predictors selected by Elastic Net. Forecast performance is assessed using MSFE, MAE, QLIKE, Model Confidence Sets, Diebold–Mariano tests, realized utility, and filtered-historical-simulation VaR and expected shortfall. The results reveal a robust horizon-dependent ranking: Markov-switching HAR performs best at short horizons, ARFIMA generally leads at the monthly horizon, and the five-day horizon is intermediate. Machine-learning models sometimes improve on HAR but do not systematically outperform the broader econometric set. These findings are robust across estimation windows, re-estimation frequencies, richer information sets, log-volatility targets, and individual equities; stock-level results are also robust to an alternative realized-volatility measure. Overall, forecast performance depends primarily on capturing the persistence and nonlinear dynamics most relevant at each horizon.

JEL Classification: C10, C50, G11, G15.
Keywords: Realized volatility, machine learning, HAR, ARFIMA, Markov switching, long memory, volatility forecasting.

*Federal Reserve Board, Washington, DC E-mail: rehim.kilic@frb.gov. The views presented in this paper are solely those of the author and do not represent those of the Board of Governors or any entities connected to the Federal Reserve System.

1 Introduction

Accurate volatility forecasts are central to asset pricing, portfolio allocation, risk measurement, margin setting, and financial-stability surveillance. Since the seminal contributions of Engle (1982), Taylor (1982) and Bollerslev (1986), a large literature has modeled conditional volatility using parametric time-series specifications. The increasing availability of high-frequency data has shifted much of this work toward realized measures, which provide observable and informative proxies for latent return variation (Andersen et al., 2001, 2003; Bauwens et al., 2012; Takahashi et al., 2023). Realized volatility nevertheless remains difficult to forecast. It is highly persistent, exhibits long-memory behavior, responds asymmetrically to adverse returns, and can change abruptly across market regimes. These features suggest that forecasting performance may depend critically on how a model represents persistence and nonlinearity.

Several modeling strategies have been developed to capture these properties. The heterogeneous autoregressive model (HAR) provides a parsimonious approximation to volatility persistence across daily, weekly, and monthly horizons. Fractionally integrated models represent long memory directly, while threshold, smooth-transition, and Markov-switching specifications allow volatility dynamics to vary across observed or latent regimes. Machine-learning (ML) methods offer a different form of flexibility: they can approximate nonlinear relationships without requiring the researcher to specify their precise functional form in advance and can accommodate rich predictor sets and complex temporal dependence. The empirical evidence on their forecasting value is mixed. ML methods improve volatility forecasts in some applications, especially when broad information sets or cross-sectional data are available, but several recent studies find that carefully estimated HAR-type specifications remain difficult to beat (Audrino et al., 2020; Christensen et al., 2023; Zhang etal. , 2023; Branco et al., 2024).

Against this background, the paper addresses a single central question: which representations of persistence and nonlinearity deliver reliable out-of-sample gains in realized-volatility forecasting, and does the generic flexibility of ML provide incremental value beyond econometric models specifically designed to capture long memory and regime dependence? This question is broader than a comparison of ML methods with the standard HAR benchmark. An ML model may improve on HAR while failing to outperform econometric models that more directly represent the relevant features of volatility dynamics. A demanding comparison must therefore evaluate ML methods against a broader econometric frontier and examine whether the relative value of the competing representations changes with the forecast horizon.

We investigate this question through a broad out-of-sample comparison centered on S&P 500 realized volatility and complemented by a cross-sectional analysis of individual-equity realized volatility. The econometric comparison set consists of the HAR model of Corsi (2009), an autoregressive fractionally integrated moving-average model (ARFIMA) (Baillie , 1996), and three

nonlinear HAR extensions: threshold HAR (THAR), smooth-transition HAR (STHAR), and Markov-switching HAR (MSHAR). These models embody distinct representations of volatility dynamics, ranging from parsimonious heterogeneous persistence and fractional long memory to observed and latent regime dependence. The ML comparison set includes XGBoost, a feedforward deep neural network, a long short-term memory network (LSTM), LSTM with attention, and a gated recurrent unit model with attention (GRU-A).

The models are evaluated under two information sets. In the baseline specification, forecasts rely exclusively on the past history of realized volatility. For the HAR-family and ML models, this history is summarized by daily, weekly, and monthly realized-volatility components, whereas ARFIMA represents the univariate history of realized volatility through its autoregressive, moving-average, and fractional-difference structure. The extended information set augments the core realized-volatility variables with daily macro-financial, uncertainty, sentiment, geopolitical-risk, factor, and return-asymmetry predictors. These auxiliary predictors are selected using a common Elastic Net protocol under which the core realized-volatility variables are always retained. Comparing the two information sets allows us to determine whether the relative forecast performance of the competing models changes when they are given access to a richer conditioning set. Throughout the paper, "baseline" and "extended" therefore refer to the restricted and expanded information sets, respectively, whereas HAR refers exclusively to the conventional heterogeneous autoregressive benchmark model.

The primary S&P 500 analysis uses a long time series and generates recursive forecasts from 2006 onward at horizons of one, five, and twenty-two trading days. Forecasts are constructed under both expanding-window estimation and rolling 10-year estimation. The evaluation period spans relatively tranquil conditions as well as major episodes of market stress, including the Global Financial Crisis, the COVID-19 shock, and the post-pandemic inflation and monetary-policy-tightening period. All models are evaluated on aligned out-of-sample observations using mean squared forecast error (MSFE), mean absolute error (MAE), and QLIKE loss. Model Confidence Sets and Diebold–Mariano tests are used to assess the statistical significance and robustness of the rankings. We also examine the economic and risk-management implications of the forecasts through realized-utility comparisons, filtered-historical-simulation Value-at-Risk (VaR) coverage tests, and joint VaR–expected-shortfall evaluation based on the FZ0 loss.

The S&P 500 analysis is complemented by a cross-sectional examination of the full set of 40 individual U.S. equities available in the VOLARE stock database. This extension asks whether the conclusions obtained for aggregate market volatility reflect features specific to an index or apply more broadly to heterogeneous firm-level volatility processes. Individual stocks provide a distinct forecasting environment because their volatility is affected by firm-specific information, idiosyncratic shocks, differences in trading activity, and greater cross-sectional heterogeneity. We apply the same model classes and broad recursive forecasting framework to

these equities over 2015–2025, using expanding and rolling 5-year estimation windows. The extended candidate predictor set also includes stock-specific changes in trading volume and the number of trades.

Because the stock histories are shorter and use a different rolling-window length from the long-sample index analysis, we additionally estimate the models for the S&P 500 over the same 2015–2025 period and under the same expanding- and rolling-window designs. This matched-sample index exercise helps distinguish genuine differences between index and firm-level volatility from effects attributable to sample length or estimation-window construction. We also compare VOLARE realized-volatility series with corresponding TAQ measures for a selected validation sample. The individual-equity analysis thus provides direct cross-sectional evidence on the generality of the index-level findings rather than serving merely as a robustness exercise based on a small number of selected firms.

The results point to a clear answer to the paper's central question. Forecast performance depends less on nonlinear flexibility in the abstract than on whether a model represents the feature of realized-volatility dynamics that is most important at the relevant forecast horizon. At the one-day horizon, MSHAR performs especially well, indicating that latent regime variation contains forecasting information not adequately captured by the linear HAR benchmark, the observed-threshold and smooth-transition specifications, or the more flexible ML architectures. At the twenty-two-day horizon, ARFIMA generally performs best, highlighting the importance of fractional long-memory persistence. The five-day horizon represents an intermediate case in which the relative performance of MSHAR and ARFIMA varies more across loss functions and estimation schemes. The central finding is therefore not simply that econometric models outperform ML models. Rather, targeted representations of latent regime dependence and fractional persistence are generally more effective than generic nonlinear flexibility, with their relative importance varying systematically across forecast horizons.

This horizon-dependent pattern is remarkably stable across expanding- and rolling-window estimation and across MSFE, MAE, and QLIKE loss. HAR remains a useful and parsimonious benchmark, but it is not uniformly on the econometric forecasting frontier once models that represent latent regimes and fractional integration are included. THAR and STHAR also allow for nonlinear dynamics, but their performance is less consistently favorable than that of MSHAR. These results demonstrate why the conclusions drawn from an ML-versus-HAR comparison can depend critically on the breadth of the econometric benchmark set.

Expanding the baseline information set to include the auxiliary predictors does not materially alter the broad ranking of the model classes. Although these predictors improve performance for some model–horizon combinations, the gains are not sufficiently systematic to allow the ML models to displace the leading econometric specifications. The relative performance of the ML methods therefore cannot be attributed solely to the breadth of the available predictor

set. Instead, the results again emphasize the importance of directly representing the persistent and regime-dependent dynamics of realized volatility.

The economic and tail-risk evidence broadly reinforces the statistical findings. Realized-utility comparisons preserve the principal horizon-dependent ranking, with MSHAR performing especially well at short horizons and ARFIMA becoming more important as the horizon increases. VaR coverage tests and joint VaR–ES comparisons provide a more calibration-oriented assessment and generate less uniform rankings, but they offer no systematic evidence that the ML specifications outperform the leading econometric models. FZ0-based Diebold–Mariano comparisons provide their clearest support for MSHAR at the short horizon.

The cross-sectional evidence shows that the main conclusions are not peculiar to S&P 500 volatility. Across the 40 individual equities, MSHAR produces the strongest overall performance at the short horizon, whereas ARFIMA becomes increasingly competitive and generally leads at the monthly horizon. ML models sometimes improve on the basic HAR benchmark, but such gains largely disappear when the comparison is made against the best model from the broader econometric set. In particular, the best-performing ML model does not systematically outperform the best econometric model across equities, horizons, estimation windows, and loss functions. The matched-sample S&P 500 results indicate that this conclusion is not simply a consequence of the shorter stock histories or the use of 5-year rolling windows.

A broad set of diagnostic and robustness exercises supports this interpretation. Quarterly rather than annual model re-estimation does not overturn the rankings, indicating that the relative performance of ML is not primarily driven by infrequent updating. The main findings also remain intact when log realized volatility is used as the target and when realized variance based on 5-minute returns is replaced by a realized-kernel measure. Neural-network training and validation diagnostics indicate that early stopping, dropout, and weight decay operate as intended, so the results cannot be attributed simply to uncontrolled overfitting. Increasing the number of random ML initializations also leaves the main conclusions unchanged.

The paper's contribution lies in the breadth and controlled nature of this comparison and in the empirical interpretation that emerges from it. Existing studies have examined stronger econometric alternatives to HAR, broader predictor sets, and various ML architectures. We build on this literature by bringing several distinct representations of persistence and nonlinearity into a common recursive forecasting design. In particular, the analysis jointly evaluates fractional integration, observed thresholds, smooth transitions, latent Markov regimes, tree-based learning, and neural-network architectures under common information sets, forecast horizons, estimation schemes, and evaluation criteria. This design makes it possible to identify not only which model performs best, but also which representation of volatility dynamics is most useful at each forecast horizon.

The paper further extends the comparison along three dimensions. First, it examines whether model rankings change when the information set is expanded through a common predictor-screening protocol. Second, it evaluates whether statistical rankings carry over to economically and operationally relevant criteria through realized utility and VaR–ES analysis. Third, the 40-equity cross-sectional analysis, together with the matched-sample S&P 500 exercise provides direct evidence on whether the index-level conclusions extend to heterogeneous firm-level volatility processes. Taken together, these elements provide a demanding assessment of whether generic ML flexibility adds forecasting value beyond interpretable econometric models designed around the principal empirical features of realized volatility.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the forecasting framework, econometric and machine-learning models, implementation protocol, out-of-sample evaluation design, and predictor-selection procedure. Section 4 describes the data and presents summary statistics for realized volatility and the predictor variables. Section 5 presents results for the S&P 500 under both the baseline and extended information sets. It also reports cross-sectional evidence for 40 individual equities and matched-sample S&P 500 comparisons using the extended information set. Section 6 concludes. Results from robustness checks—including quarterly re-estimation, alternative realized-volatility transformations and measures, and the sensitivity of ML results to the number of seeds used for ensembling—are reported in Appendix A. Appendices B–E provide further details on the forecasting models, statistical tests, implementation and hyperparameter choices, data and predictors, sample coverage, and equity summary statistics. They also report ML feature importance, illustrative econometric parameter estimates, and neural-network training and regularization diagnostics. An extended Online Appendix presents additional supporting results, including a comparison of the relatively new VOLARE series with realized-volatility measures constructed from NYSE Trade and Quote (TAQ) data.

2 Related Literature

This paper relates to three strands of research: econometric modeling of realized-volatility dynamics, machine-learning applications to volatility forecasting, and comparative studies that evaluate whether flexible ML methods provide incremental forecasting value beyond conventional and nonlinear econometric models. Across these strands, the central issue for our analysis is how alternative models represent the persistence and nonlinearity of realized volatility and whether the usefulness of those representations changes with the forecast horizon.

The first strand builds on the use of high-frequency returns to construct ex post measures of return variation for forecasting and risk management. A prominent empirical feature of realized volatility is its strong persistence, often characterized as long-memory behavior (Andersen et

al. , 2001). Two econometric approaches have become especially important for representing this persistence. ARFIMA models explicitly parameterize fractional integration and the resulting slow decay of volatility shocks (Baillie , 1996; Andersen et al. , 2003). The heterogeneous autoregressive model of Corsi (2009), by contrast, provides a parsimonious approximation to persistent volatility dynamics through daily, weekly, and monthly realized-volatility components. Its simplicity, interpretability, and forecasting performance have made HAR the conventional benchmark in much of the realized-volatility literature. Subsequent HAR extensions incorporate asymmetry, signed variation, jumps, and macro-financial information (Corsi and Renó , 2012; Patton and Sheppard , 2015; Izzeldin et al. , 2019).

Although HAR and ARFIMA can both reproduce highly persistent dynamics, they embody different representations of that persistence. HAR approximates long memory through a finite collection of heterogeneous volatility components, whereas ARFIMA represents fractional persistence directly. Evidence that fractional integration can remain relevant after controlling for HAR-type dynamics suggests that the two approaches need not be empirically interchangeable (Baillie et al. , 2019). This distinction is particularly relevant in a multi-horizon forecasting exercise: the parsimonious HAR approximation may be adequate at some horizons, while the slower decay implied by fractional integration may become more important at others.

Persistence is not the only salient feature of realized volatility. Volatility dynamics can also vary with market conditions, respond asymmetrically to shocks, and change abruptly during periods of financial stress. Threshold and smooth-transition models represent such variation as a function of an observed transition variable, with the former allowing discrete changes and the latter allowing gradual adjustment. Markov-switching models instead treat the prevailing volatility state as latent and probabilistic. These specifications consequently capture different forms of nonlinearity rather than merely providing interchangeable extensions of HAR. Their inclusion allows the forecast comparison to assess whether observed thresholds, gradual transitions, or latent regime changes provide useful information beyond the linear HAR benchmark and whether their performance differs across forecast horizons.

A second strand of the literature applies ML methods to volatility forecasting. The methods considered include regularized regressions such as LASSO and Elastic Net, boosted trees, feedforward neural networks, and recurrent architectures such as LSTM. These methods can accommodate nonlinear interactions, large information sets, and complex temporal relationships without requiring the researcher to specify a particular form of regime dependence in advance. Their empirical performance, however, varies substantially across applications. Some studies emphasize low-frequency macro-financial predictors (Mittnik et al. , 2015; Bucci , 2020), while others exploit firm-level, panel, or cross-sectional information (Christensen et al. , 2023; Zhang etal. , 2023), or large predictor sets constructed from order-book and news information

(Rahimika and Poon , 2024). These differences in information sets, sampling frequency, forecast targets, and evaluation designs help explain why the reported gains from ML are not uniform.

The available evidence suggests that ML improvements are more likely when the forecasting problem contains rich predictors, nonlinear interactions, or cross-sectional information that cannot be summarized adequately by a small univariate model. In more conventional realized-volatility forecasting settings, however, carefully estimated HAR-type specifications often remain highly competitive (Audrino et al., 2020; Branco et al., 2024). This evidence raises an important interpretive issue. Failure to outperform HAR does not necessarily imply that nonlinear modeling is unhelpful, just as outperforming HAR does not establish that generic ML flexibility is superior to more targeted econometric nonlinearities. The appropriate conclusion depends on the breadth of the econometric comparison set.

The third and most directly related strand therefore concerns the benchmarks against which ML methods are evaluated. HAR is a natural reference model because of its established forecasting performance, and it features prominently in many comparative studies. Nevertheless, comparisons centered primarily on HAR-class of models may not reveal whether apparent ML gains arise from machine learning specifically or from the ability to capture dynamics omitted by linear HAR. ARFIMA provides a direct representation of fractional persistence, while threshold, smooth-transition, and Markov-switching models offer interpretable representations of observed or latent regime dependence. A demanding assessment of ML forecasting value should therefore determine whether ML methods outperform not only HAR, but also econometric models designed to capture these features directly.

Several studies move in this broader direction, so our analysis should be viewed as extending rather than initiating such comparisons. Bucci (2020), for example, compares neural-network forecasts with a smooth-transition autoregressive specification, although the analysis is conducted at monthly frequency and does not consider the broader collection of nonlinear HAR models examined here. Other studies evaluate neural networks under wider information sets, while recent work such as Branco et al. (2024) provides extensive comparisons involving HAR-type benchmarks. These contributions demonstrate that both the information set and the econometric benchmark matter for conclusions about ML performance. They also motivate a comparison in which fractional integration, observed thresholds, smooth transitions, latent regimes, tree-based learning, and neural-network architectures are evaluated under a common forecasting protocol.

Our analysis builds on these literatures in three related ways. First, it places alternative representations of persistence and nonlinearity within the same recursive out-of-sample design. The comparison includes HAR, ARFIMA, THAR, STHAR, and MSHAR alongside XGBoost, feedforward neural networks, and recurrent neural-network architectures. This broader model set permits a more precise question than whether ML beats HAR: it allows us to examine which

representation of volatility dynamics is most useful at each forecast horizon and whether generic ML flexibility provides additional value beyond targeted econometric structures.

Second, we evaluate the competing models under both a baseline information set restricted to the past history of realized volatility and an extended-information set constructed using a common Elastic Net screening protocol. This design helps separate gains attributable to additional conditioning information from those attributable to the forecasting model itself. The statistical comparison is supplemented by realized-utility and VaR–expected-shortfall evaluations, which assess whether differences in forecast loss carry over to economic and risk-management applications.

Third, we examine whether the conclusions obtained for S&P 500 realized volatility extend to the full set of 40 individual U.S. equities available in the VOLARE stock database. Existing studies using firm-level and panel information show that cross-sectional data can create a setting in which ML methods may be particularly valuable (Christensen et al., 2023; Zhang etal. , 2023). Our stock-level analysis addresses a related but distinct question: whether the relative usefulness of latent regime dependence, fractional persistence, and ML flexibility generalizes across heterogeneous firm-level volatility processes when the competing models are evaluated asset by asset under a common design. Matched-sample S&P 500 results help distinguish index–equity differences from the effects of sample length and estimation-window construction. The resulting framework connects the econometric and ML literatures by focusing on the economic structure represented by each model, rather than treating the comparison solely as a contest between model classes.

3 Empirical Methodology

This section describes the forecasting framework and the models used in the empirical comparison. The organization follows the economic structure of the forecasting problem. Linear HAR and ARFIMA models capture persistence and long memory. THAR and STHAR introduce observable state dependence through threshold and smooth-transition mechanisms. MSHAR allows latent regimes to govern the volatility process through a Markov chain. The ML models provide flexible nonlinear alternatives based on boosted trees, feedforward neural networks, and recurrent neural networks. Throughout, the purpose is not simply to compare "econometrics" with "machine learning," but to identify which representation of persistence, nonlinearity, and information processing is most useful for realized-volatility forecasting.

3.1 Forecast setup and common notation

Let RV_t denote the daily realized variance computed from intraday returns on trading day t. Under standard assumptions, RV_t estimates the integrated variance

$$IV_t = \int_{t-1}^{t} \sigma_s^2 \, ds, \tag{1}$$

where σ_s is the instantaneous volatility. With n intraday log returns r_s, realized variance is

$$RV_t = \sum_{s=1}^{n} r_s^2, \qquad r_s = \log\left(\frac{P_s}{P_{s-1}}\right), \tag{2}$$

where P_s and P_{s-1} are transaction or quote prices at adjacent intraday sampling points.[1] Following the realized-volatility literature, we use five-minute returns, which balance consistency of the realized-variance estimator against market-microstructure noise (Andersen et al. , 2001; Barndorff et al. , 2002; Liu et al. , 2015).[2]

The primary forecasting target is daily realized volatility in percentage points,

$$y_t = 100\sqrt{RV_t}. \tag{3}$$

For each horizon $h \in \{1, 5, 22\}$, the statistical forecasting exercise predicts y_{t+h} using information available at date t. These horizons correspond to approximately one trading day, one trading week, and one trading month. All statistical forecast losses are computed on the percentage-volatility scale in (3); when a model is estimated on a transformed scale, its forecasts are mapped back to the level scale before evaluation.

3.2 Forecasting models

This subsection summarizes the forecasting specifications used in the empirical comparison. The objective is not to provide a general methodological review of HAR, ARFIMA, regime-switching regressions, or machine-learning algorithms, but to define the implemented models and to make the comparison across model classes transparent. All models are estimated recursively and produce direct forecasts of the same horizon-specific target y_{t+h}, where $h \in \{1, 5, 22\}$. The

[1] Note that s indexes intraday observations within day t.

[2] Liu et al. (2015) show that realized variance constructed from five-minute returns performs well relative to measures based on alternative sampling frequencies. Nevertheless, to assess whether the results depend on the particular measurement of realized volatility, we repeat the forecasting analysis using the realized kernel, which is designed to be robust to market microstructure noise. The resulting model rankings and principal conclusions remain broadly unchanged (see, Appendix A.3), indicating that the findings are not driven by the use of the five-minute realized-volatility measure.

forecast origin is date t. Hence, all predictors are observed at or before t, while the realization being forecast is dated $t+h$.

This timing convention is especially important for the HAR variables. The HAR components are computed using realized-volatility information available through the forecast origin t: the daily component is y_t, the weekly component is the average of the most recent five daily observations ending at t, and the monthly component is the corresponding average over the most recent twenty-two observations ending at t. Thus, the HAR variables are contemporaneous with the forecast origin, but they are predetermined relative to the forecast target y_{t+h}. Additional predictors entering the HAR-X specifications are dated analogously so that they are available at the forecast origin. This preserves the real-time nature of the forecasting exercise.

The model set is organized around three empirical mechanisms that are central to realized-volatility forecasting: multi-scale persistence, nonlinear or state-dependent dynamics, and flexible nonlinear approximation. The first group contains HAR and ARFIMA models. The second group contains nonlinear HAR extensions with observed or latent regimes: threshold HAR (THAR), smooth-transition HAR (STHAR), and Markov-switching HAR (MSHAR). The third group contains machine-learning specifications: XGBoost, a feedforward neural network, LSTM, LSTM with attention, and GRU with attention. Detailed estimation algorithms, transition-variable screening procedures, neural-network architecture choices, and hyperparameter grids are reported in Appendix B.

3.2.1 HAR and ARFIMA

The benchmark HAR model of Corsi (2009) approximates the strong persistence of realized volatility through daily, weekly, and monthly realized-volatility components:

$$y_{t+h} = \alpha + \beta_1 \overline{y}_{t,1} + \beta_5 \overline{y}_{t,5} + \beta_{22} \overline{y}_{t,22} + \varepsilon_{t+h}, \tag{4}$$

where

$$\overline{y}_{t,w} = w^{-1} \sum_{j=1}^{w} y_{t-j+1}, \qquad w \in \{1, 5, 22\}.$$

The HAR specification is estimated by OLS and serves as the principal linear benchmark. In the extended HAR-X specification, the HAR components are forced into the model and the remaining candidate predictors are selected using the common Elastic Net screening protocol described in Section 3.5.

The ARFIMA model provides an explicit fractional-integration alternative to the HAR approximation of long memory (Baillie , 1996; Andersen et al. , 2003). We use the parsimonious

ARFIMA$(0, d, 1)$ specification throughout the reported analysis:

$$(1 - L)^d (y_{t+h} - \mu) = (1 + \theta L)\varepsilon_{t+h}, \qquad \varepsilon_{t+h} \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

with d denoting the fractional-memory parameter. The corresponding ARFIMAX specification allows EN-selected predictors to enter the conditional mean before fractional differencing. The purpose of including ARFIMA is to test whether explicitly parameterizing long memory improves on the HAR approximation, particularly at longer forecast horizons.[3]

3.2.2 Observable and latent regime dependence

The nonlinear econometric specifications are designed to assess whether realized volatility forecasts improve when the HAR relation is allowed to vary across states. The comparison is disciplined by imposing the same predictor structure across the nonlinear HAR models. The regime-switching block contains only the intercept and the daily HAR component,

$$\mathbf{x}_t^{\text{sw}} = (1, \overline{y}_{t,1})',$$

while the weekly and monthly HAR components, together with any EN-selected HAR-X predictors, enter through a common-coefficient block $\mathbf{x}_t^{\text{cm}}$. This restriction allows the nonlinear models to differ in how states are defined, while holding fixed the set of predictors that can switch across states.

THAR defines regimes using an observed transition variable and an estimated threshold. In the two-regime case used in the empirical analysis,

$$y_{t+h} = \mathbf{1}(z_t \le c)\, \boldsymbol{\beta}_1' \mathbf{x}_t^{\text{sw}} + \mathbf{1}(z_t > c)\, \boldsymbol{\beta}_2' \mathbf{x}_t^{\text{sw}} + \boldsymbol{\delta}' \mathbf{x}_t^{\text{cm}} + \varepsilon_{t+h}. \tag{6}$$

STHAR replaces the discrete threshold with a smooth logistic transition:

$$y_{t+h} = \boldsymbol{\phi}' \mathbf{x}_t^{\text{cm}} + \boldsymbol{\beta}' \mathbf{x}_t^{\text{sw}} + G(z_t; \gamma, c)\, \boldsymbol{\delta}' \mathbf{x}_t^{\text{sw}} + \varepsilon_{t+h}, \qquad G(z_t; \gamma, c) = [1 + \exp\{-\gamma(z_t - c)\}]^{-1}. \tag{7}$$

Both models use the same recursively selected pool of RV-based transition candidates, so their performance differences reflect the form of observable state dependence—sharp versus smooth transitions—rather than different information sets.

[3] ARFIMA serves as a linear long-memory benchmark alongside HAR. We do not construct nonlinear extensions of ARFIMA: unlike the HAR components, the fractional differencing operator does not decompose naturally into a "switching" block, so threshold or smooth-transition ARFIMA models would require switching on d itself or on the ARMA innovations — an approach that is computationally demanding and seldom used in realized volatility forecasting (see Baillie , 1996, for the standard treatment) and Kılıç (2011) for a nonlinear extension to conditional variance. The nonlinear extensions are therefore applied exclusively to the HAR framework.

MSHAR allows the regimes to be latent rather than determined by an observed transition variable. The two-state specification is

$$y_{t+h} = \boldsymbol{\beta}'_{s_t}\mathbf{x}_t^{\mathrm{sw}} + \boldsymbol{\delta}'\mathbf{x}_t^{\mathrm{cm}} + \varepsilon_{t+h}, \qquad \varepsilon_{t+h} \mid s_t \sim \mathcal{N}(0, \sigma^2_{s_t}), \tag{8}$$

where $s_t \in \{1, 2\}$ follows a first-order Markov chain. The out-of-sample forecast is the regime-probability-weighted conditional mean computed from filtered probabilities using only information available through the forecast origin:

$$\hat{y}_{t+h} = \sum_{r=1}^{2} \hat{P}(s_t = r \mid \mathcal{F}_t; \hat{\boldsymbol{\theta}})\hat{\boldsymbol{\beta}}'_r\mathbf{x}_t^{\mathrm{sw}} + \hat{\boldsymbol{\delta}}'\mathbf{x}_t^{\mathrm{cm}}. \tag{9}$$

The MSHAR specification therefore tests whether latent regime dependence is more useful than observed threshold or smooth-transition mechanisms for forecasting realized volatility.

3.2.3 Machine-learning models

The machine-learning specifications provide flexible nonlinear benchmarks that do not impose the HAR, fractional-integration, or regime-switching structure used by the econometric models. Each ML model uses the same forecast target, forecast origins, and candidate information set as the corresponding HAR-X specification. Predictors are standardized within each training window; hyperparameters are selected using the time-series validation procedure described in Section B.7.1; and neural-network models are trained with validation-based early stopping.

XGBoost represents the forecast as an additive ensemble of regression trees,

$$\hat{y}_{t+h} = \sum_{k=1}^{K} f_k(\mathbf{x}_t), \qquad f_k \in \mathcal{F}, \tag{10}$$

where $\mathcal{F}$ is the space of regression trees. This specification captures nonlinearities and interactions among predictors while regularizing tree depth, leaf weights, and minimum leaf size.

The feedforward neural network maps the standardized feature vector $\mathbf{x}_t$ into a scalar forecast through a single nonlinear hidden layer:

$$\hat{y}_{t+h} = \mathbf{w}'_2\mathrm{Drop}_p\left[\sigma_\alpha(\mathbf{W}_1\mathbf{x}_t + \mathbf{b}_1)\right] + b_2, \tag{11}$$

where $\sigma(.)$ is the logistic activation and $\mathrm{Drop}_p(\cdot)$ denotes dropout regularization. The recurrent specifications—LSTM, LSTM with additive attention, and GRU with additive attention—process the recent input history as a sequence rather than as a static feature vector. The attention variants form a data-driven weighted average of hidden states before the final output

layer, allowing the model to emphasize different parts of the recent history when forming the volatility forecast.

For each forecast origin, the final neural-network and XGBoost forecasts are computed using parameter estimates and preprocessing transformations learned only from the corresponding training window. Forecasts from independently seeded neural-network fits are averaged to reduce sensitivity to initialization. This common protocol ensures that differences in forecast performance can be interpreted as differences across model classes rather than differences in forecast timing, predictor availability, or evaluation samples. Additional details on ML models and implementation are provided in Appendix B.

3.3 Out-of-Sample Evaluation Protocol

3.3.1 Recursive forecasting design

All models are evaluated in a recursive out-of-sample framework. For the aggregate S&P 500 series, the initial estimation sample covers 1996–2005 and the out-of-sample evaluation period begins in 2006. At each re-estimation date, model estimation, predictor screening, hyperparameter tuning, standardization, validation, and early stopping use only observations contained in the relevant training window. The fitted models then produce daily forecasts until the next scheduled re-estimation date.

At each re-estimation date T, the horizon-h estimation sample includes only forecast-origin observations $s < T$ whose associated target y_{s+h} has been realized before the new evaluation period begins. This horizon-specific label-availability restriction ensures that no dependent-variable realization from the subsequent out-of-sample period enters model estimation, predictor screening, standardization, validation, or tuning.

Two estimation-window schemes are considered. The expanding-window scheme always starts in 1996 and grows through time, whereas the rolling-window scheme uses the most recent ten years of observations. The expanding design emphasizes estimation precision, while the rolling design permits the model parameters to adapt more quickly to structural change.

For each forecast horizon $h \in \{1, 5, 22\}$, the models use information available at forecast origin t to predict the daily realized volatility observed at $t+h$, denoted y_{t+h}. The out-of-sample observations are indexed and assigned to calendar years according to the forecast-origin date t, rather than the target date $t + h$. Thus, a forecast formed near the end of a calendar year remains part of that year's evaluation sample even when its realization occurs in the following year. The horizon-specific target is constructed over the continuous time series rather than separately within each calendar year, so forecasts are not discarded merely because the target date crosses a year boundary.

Consequently, the number of forecasts in year Y is

$$N_Y^{(h)} = \#\left\{t : \ t \in Y, \ X_t \text{ and } y_{t+h} \text{ are available}\right\},$$

where X_t denotes the information set available at the forecast origin. When the same set of forecast-origin dates is available and the corresponding future realizations exist, $N_Y^{(h)}$ is identical across $h = 1, 5$, and 22. It is therefore not generally equal to $n_Y - h$, where n_Y is the number of observations in calendar year Y. Such a reduction would arise only if the target were constructed separately within each year, if both t and $t + h$ were required to fall in the same year, or near the terminal date of the full sample when future target observations are unavailable. Detailed estimation-sample sizes and annual out-of-sample forecast counts for the stock and S&P 500 exercises are reported in Appendix C.3. The reported sample sizes are the actual numbers of aligned and evaluable forecast-origin observations after applying the relevant target-availability and complete-case requirements.

The individual-equity exercise follows the same forecasting and indexing conventions but uses asset-specific samples and a shorter five-year rolling window. This adjustment reflects the shorter realized-volatility histories available for individual stocks and gives the models greater flexibility to adapt to firm-level volatility dynamics. In all cases, models are estimated separately by asset, forecast horizon, and window type.

The baseline implementation re-estimates the models once per year. A separate robustness exercise uses quarterly re-estimation under the same rolling-window design and the same Elastic Net predictor-screening protocol. This comparison isolates whether the main model rankings are sensitive to the frequency with which the forecasting models are updated.

3.3.2 Forecast accuracy metrics

Forecast accuracy is assessed using three loss functions, all evaluated on the level of realized volatility in percentage points:

$$\text{MSFE} = T^{-1} \sum_t \left(y_{t+h} - \hat{y}_{t+h}\right)^2, \tag{12}$$

$$\text{MAE} = T^{-1} \sum_t \left|y_{t+h} - \hat{y}_{t+h}\right|, \tag{13}$$

$$\text{QLIKE} = T^{-1} \sum_t \left[\log \hat{y}_{t+h} + \frac{y_{t+h}}{\hat{y}_{t+h}}\right]. \tag{14}$$

MSFE emphasizes large errors, MAE summarizes average absolute forecast errors, and QLIKE is a quasi-likelihood loss that is commonly used in volatility forecasting because it is robust to noisy volatility proxies and penalizes underprediction relatively heavily (Patton , 2011).

Pairwise comparisons are conducted with Diebold–Mariano tests (Diebold and Mariano , 1995) using horizon-appropriate long-run variance corrections.[4] Model Confidence Sets are computed at the 5% level (Hansen et al. , 2011) to identify models that cannot be statistically distinguished from the best-performing specification.

3.3.3 Realized Utility

To complement statistical losses, we evaluate the economic value of volatility forecasts using the realized-utility criterion of Bollerslev et al. (2018). The exercise considers a mean–variance investor who scales risky-asset exposure using the model-implied volatility forecast. Following the literature, we assume a constant conditional Sharpe ratio $SR = 0.40$ and relative risk aversion $\gamma = 2$, implying a target volatility of $SR/\gamma = 0.20$. Let $\widehat{RV}_{t,h}^{(m)}$ denote model m's horizon-h variance forecast for target date t, and let $RV_{t,h}$ denote the corresponding realized variance. Average realized utility is

$$\widehat{RU}_h^{(m)} = \frac{1}{T_{\text{test}}} \sum_{t=1}^{T_{\text{test}}} \left(0.08 \sqrt{\frac{RV_{t,h}}{\widehat{RV}_{t,h}^{(m)}}} - 0.04 \frac{RV_{t,h}}{\widehat{RV}_{t,h}^{(m)}} \right). \tag{15}$$

A perfect forecast satisfies $\widehat{RV}_{t,h}^{(m)} = RV_{t,h}$ and yields $\widehat{RU}_h^{(m)} = 0.04$, or 4 percent of wealth. Higher values therefore indicate greater economic value from using model m in volatility-targeted allocation.

3.4 VaR, Expected Shortfall, and FZ0-Based Forecast Comparison.

To evaluate tail-risk performance, we construct horizon-h Value-at-Risk (VaR) and Expected Shortfall (ES) forecasts using filtered historical simulation (FHS). Let $r_{t,h}$ denote the cumulative log return over horizon h ending at target date t:

$$r_{t,h} = \sum_{j=0}^{h-1} r_{t-j}, \qquad h \in \{1, 5, 22\}.$$

For model m, the volatility forecast is converted into return units as

$$\hat{\sigma}_{t,h}^{(m)} = \frac{\widehat{y}_{t,h}^{(m)}}{100}.$$

[4] For each horizon h, the Diebold–Mariano test uses a Bartlett long-run variance estimator based on autocovariances through lag $h - 1$ and applies the Harvey–Leybourne–Newbold small-sample correction.

The model-specific standardized return innovation is

$$\hat{z}_{t,h}^{(m)} = \frac{r_{t,h}}{\hat{\sigma}_{t,h}^{(m)}}. \tag{16}$$

For a forecast targeting date d, the historical innovation pool uses only observations available through $d-h$,

$$\mathcal{Z}_{d,h}^{(m)} = \left\{ \hat{z}_{t,h}^{(m)} : t \le d-h \right\}.$$

The VaR forecast at tail probability α is the empirical α-quantile of the rescaled distribution,

$$\widehat{VaR}_{d,h}^{(m,\alpha)} = Q_\alpha \left(\hat{\sigma}_{d,h}^{(m)} \, \mathcal{Z}_{d,h}^{(m)} \right), \tag{17}$$

and the ES forecast is the average of simulated returns below the VaR cutoff,

$$\widehat{ES}_{d,h}^{(m,\alpha)} = \mathbb{E}\left[\hat{\sigma}_{d,h}^{(m)} \hat{z}_{t,h}^{(m)} \;\middle|\; \hat{\sigma}_{d,h}^{(m)} \hat{z}_{t,h}^{(m)} \le \widehat{VaR}_{d,h}^{(m,\alpha)} \right]. \tag{18}$$

VaR calibration is assessed with the exception indicator

$$I_{d,h}^{(m,\alpha)} = \mathbf{1}\left\{ r_{d,h} < \widehat{VaR}_{d,h}^{(m,\alpha)} \right\}. \tag{19}$$

We report exceedance rates and apply the unconditional- and conditional-coverage tests of Kupiec (1995) and Christoffersen (1998). For $h > 1$, overlapping cumulative returns induce mechanical serial dependence in adjacent exceptions, so formal coverage tests are applied to non-overlapping target dates, while full-sample exceedance rates are reported for transparency.

Joint VaR–ES accuracy is evaluated with the FZ0 loss of Fissler and Ziegel (2016), following Patton et al. (2019). Under the left-tail convention $\widehat{ES}_{d,h}^{(m,\alpha)} \le \widehat{VaR}_{d,h}^{(m,\alpha)} < 0$, define $v_{d,h}^{(m,\alpha)} = \widehat{VaR}_{d,h}^{(m,\alpha)}$ and $e_{d,h}^{(m,\alpha)} = \widehat{ES}_{d,h}^{(m,\alpha)}$. The period loss is

$$\begin{aligned} L_{FZ0}(r_{d,h}, v_{d,h}^{(m,\alpha)}, e_{d,h}^{(m,\alpha)}; \alpha) = & -\frac{1}{\alpha e_{d,h}^{(m,\alpha)}} \mathbf{1}\{r_{d,h} \le v_{d,h}^{(m,\alpha)}\}\left(v_{d,h}^{(m,\alpha)} - r_{d,h}\right) \\ & + \frac{v_{d,h}^{(m,\alpha)}}{e_{d,h}^{(m,\alpha)}} + \log\!\left(-e_{d,h}^{(m,\alpha)}\right) - 1. \end{aligned} \tag{20}$$

This score is strictly consistent for the pair (VaR_α, ES_α), so lower average values indicate more accurate joint tail-risk forecasts. Pairwise comparisons use Diebold–Mariano tests applied to the FZ0 loss differential

$$\Delta L_{d,h}^{ij} = L_{FZ0}^{(i)}(d,h) - L_{FZ0}^{(j)}(d,h),$$

with null hypothesis

$$H_0 : \mathbb{E}\Big[\Delta L_{d,h}^{ij}\Big] = 0, \qquad H_1 : \mathbb{E}\Big[\Delta L_{d,h}^{ij}\Big] \neq 0. \tag{21}$$

Negative values in the upper-triangular DM tables favor the column model, while positive values favor the row model.

3.5 Shared Predictor-Selection Protocol

Each model is evaluated in two variants. The baseline variant uses only the model's core realized-volatility predictors: the HAR block for HAR-type models, the relevant ARMA/fractional-difference structure for ARFIMA, and the corresponding HAR variables supplied to the ML models. The extended variant augments the core predictors with macro-financial, uncertainty, sentiment, geopolitical-risk, factor, and return-asymmetry variables. The candidate set includes $\log(VIX_t)$; Fama–French factors; short- and long-term interest-rate variables; EPU (Baker et al., 2016); the Composite Indicator of Systemic Stress; the ADS business-conditions index (Aruba etal., 2009); the San Francisco Fed news sentiment index (Buckman et al., 2020); a geopolitical-risk index; and short-horizon negative-return averages. Variable definitions and sources are reported in Section 4 and Appendix C.

The auxiliary predictor set is intentionally broad and contains correlated measures of financial conditions, uncertainty, and market stress. To avoid unstable recursive estimates and to keep the effective information set comparable across model classes, we screen the auxiliary predictors with Elastic Net regularization (Zou and Hastie , 2005). The penalty applied to the auxiliary coefficient vector $\boldsymbol{\gamma}$ is

$$\mathcal{P}(\boldsymbol{\gamma}; \rho, \lambda) = \lambda \left[\frac{1-\rho}{2} \|\boldsymbol{\gamma}\|_2^2 + \rho \|\boldsymbol{\gamma}\|_1 \right], \tag{22}$$

where $\rho \in [0, 1]$ controls the Lasso–Ridge mix and $\lambda > 0$ controls the overall amount of shrinkage.

The core realized-volatility regressors are forced in and are not penalized. Operationally, the target and each auxiliary predictor are first residualized with respect to the forced-in block. Elastic Net is then applied to the residualized auxiliary predictors, and the final specification is the union of the forced-in block and the selected auxiliary variables. This residualization approach ensures that the screening step selects predictors for their incremental contribution beyond the model's core volatility dynamics rather than for their ability to proxy the HAR terms.

The Elastic Net hyperparameters (λ, ρ) are selected by five-fold time-series cross-validation. The mixing parameter is searched over $\{0.10, 0.30, 0.50, 0.70, 0.90, 0.95, 1.00\}$. A predictor is retained only if it has a nonzero coefficient in the full-sample Elastic Net fit for that training

window and is selected in at least three of the five cross-validation folds at the selected hyperparameters. This stability screen reduces the likelihood that a predictor enters the recursive forecasting exercise because of a transient or fold-specific correlation.

4 Data & Summary Statistics

The realized-volatility series are constructed from cleaned high-frequency price data. For the individual stocks, we use the VOLARE database, which provides standardized realized measures computed from ultra-high-frequency data using a documented asset-level pipeline that accounts for trading calendars, timestamp precision, irregular observations, microstructure noise, and outlier filtering (Cipollini et al., 2026). VOLARE obtains raw high-frequency data from Kibot and applies the high-frequency price-cleaning procedure of Brownlees and Gallo (2006), including outlier detection before regular sampling and realized-measure construction. The S&P 500 benchmark series is constructed and maintained by the Research and Statistics Division (R&S) of the Federal Reserve Board (FRB) using raw tick-by-tick data for the S&P 500 index obtained from LSEG Tick History (formerly Thomson Reuters Tick History). In constructing this series, prices are screened for outliers, restricted to regular trading hours, and sampled on a 5-minute grid, producing 78 intraday intervals on a full trading day. Daily realized variance is then computed as the sum of squared 5-minute log returns, with days containing missing or irregular intraday observations adjusted according to the available valid grid returns. Because intraday realized variance captures only variation during market hours, the open-to-close measure is rescaled by the ratio of the unconditional variance of daily close-to-close returns to the average open-to-close realized variance, thereby incorporating an adjustment for overnight volatility. The forecasting target throughout the paper is daily percentage realized volatility, $y_t = 100\sqrt{RV_t}$. Thus, the S&P 500 and stock-level analyses use comparable five-minute realized-volatility measures, with the stock series obtained from VOLARE and the S&P 500 series constructed by Federal Reserve Board Research and Statistics using the same cleaning, sampling, and realized-variance logic.

The long S&P 500 sample runs from 1996-01-02 through 2025-11-25 and is used for the main index-level forecasting analysis. For the individual-equity extension, we use the full set of 40 U.S. stocks available in the VOLARE high-frequency database (Cipollini et al., 2026). Thus, the equities were not selected ex post on the basis of their volatility characteristics or the forecasting performance of the competing models. The stock-level realized-volatility measure is constructed from the 5-minute realized variance variable, $rv5$, and transformed as $100\sqrt{rv5_t}$. The matched stock sample covers 2015-01-02 through 2025-11-25. A matched-sample S&P 500

series is reported alongside the stock results to facilitate comparison, but it is not included in the 40-stock cross-sectional summaries.[5]

The predictor timing follows the forecast-origin convention used throughout the paper. For a forecast made at date t, the HAR variables are computed using realized-volatility observations available through t: the daily component is y_t, while the weekly and monthly components are rolling averages ending at t. The extended predictors are likewise dated so that they are available at the forecast origin. Thus, the regressors are predetermined relative to the forecast target y_{t+h} for $h \in \{1, 5, 22\}$.[6] Detailed definitions, transformations, and sources for all predictors are provided in Appendix C.

Table 1 summarizes the realized-volatility and predictor data used in the forecasting exercises. Panel A reports the long-sample S&P 500 variables. The realized-volatility series is strongly right-skewed and leptokurtic, consistent with infrequent but large volatility spikes. The HAR components are smoother than the daily series, but retain substantial dispersion and tail behavior, reflecting the persistence of realized volatility. The extended variables cover return asymmetry, option-implied volatility, macro-financial conditions, sentiment, uncertainty, systemic stress, geopolitical risk, and Fama–French equity factors. Panel B summarizes the 40-stock cross section over the matched 2015–2025 sample. Individual-equity realized volatility is higher on average than the matched S&P 500 series and displays substantial cross-sectional heterogeneity. The stock-specific activity variables—daily changes in log trading volume and log number of trades—are centered close to zero but have meaningful variation, which is consistent with their role as proxies for firm-level information arrival and trading intensity. Ticker-level descriptive statistics are reported in Appendix C.

Figure 1 reports sample autocorrelation functions for daily percentage realized volatility. The figure plots the S&P 500 autocorrelation function together with the lag-by-lag minimum, maximum, and average autocorrelations across the 40 stocks; the individual stock autocorrelations are shown in the background. Realized volatility is highly persistent for both the market index and individual stocks. The S&P 500 autocorrelation function decays more slowly than

[5]The 40-stock sample should not be interpreted as a random or representative sample of the U.S. equity universe. Although we use all U.S. equities available in VOLARE rather than selecting firms according to their realized-volatility behavior or forecast results, inclusion in the database and the availability of a consistent high-frequency history over 2015–2025 may favor relatively large, liquid, and continuously traded firms. Requiring a balanced sample may therefore introduce data-availability or survivorship-related selection. Accordingly, the stock-level evidence is interpreted as showing whether the index-level findings extend across this set of continuously observed equities, rather than as establishing their applicability to all U.S. stocks, particularly smaller, less liquid, newly listed, or delisted firms.

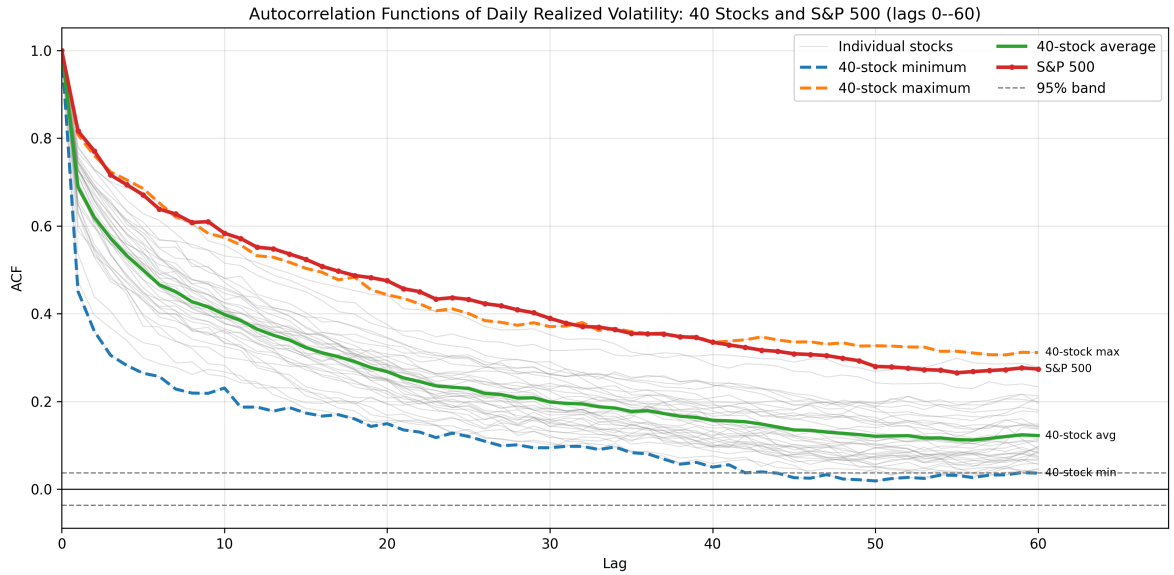
[6]All predictors in the extended information set are observed at daily frequency; the analysis therefore does not combine daily realized volatility with lower-frequency macroeconomic or survey data. Forecasts are formed after the close of trading on date t, when y_t and the date-t HAR-X predictors are observable. Accordingly, no date-$t+1$ or later information is used to forecast y_{t+h}. We should also note that final-vintage daily series are used where real-time vintages are unavailable and hence, the exercises with expanded information set should be interpreted as pseudo-real-time with respect to foirecast-origin alignment, not as a full real-time-vintage evaluation.

Table 1: Summary statistics for realized volatility and predictors

Variable	Mean	Median	Std. dev.	Min.	Max.	Skew.	Ex. kurt.
Panel A. S&P 500 long sample (1996-2025) realized volatility & predictors							
Real Vol(y_t)	1.001	0.840	0.673	0.157	11.017	3.610	24.786
HAR weekly ($\bar{y}_{t,5}$)	1.001	0.849	0.608	0.232	6.765	3.192	17.357
HAR monthly($\bar{y}_{t,22}$)	1.002	0.870	0.546	0.280	5.144	2.838	13.193
Downside return daily $r^-_{d,t}$	-0.004	0.000	0.008	-0.128	0.000	-4.126	29.960
Downside return weekly $r^-_{w,t}$	-0.002	0.000	0.003	-0.040	0.000	-3.794	23.334
Downside return monthly $r^-_{m,t}$	-0.001	0.000	0.001	-0.019	0.000	-4.101	25.938
$\Delta T3M$	-0.000	0.000	0.046	-0.810	0.760	-0.506	58.766
$T10Y3M$	1.359	1.410	1.254	-1.780	3.850	-0.132	-0.649
$\log(VIX_t)$	2.942	2.925	0.348	2.213	4.415	0.567	0.445
ADS Index (ADS_t)	-0.174	-0.057	1.782	-26.681	9.640	-8.505	120.097
$\log(EPU_t)$	4.471	4.464	0.714	1.200	6.934	-0.040	0.422
News sentiment ($NEWS_t$)	0.023	0.040	0.191	-0.670	0.450	-0.554	0.150
$\log(CISS_t)$	0.097	0.036	0.143	0.000	0.897	2.770	9.114
$\log(GPR_t^{AI})$	4.582	4.575	0.387	3.195	6.238	0.109	0.271
Rm_Rf_t	0.039	0.080	1.220	-12.010	11.360	-0.202	9.179
SMB_t	0.003	0.010	0.642	-4.580	5.710	0.077	3.745
HML_t	0.006	-0.010	0.757	-5.030	6.730	0.274	6.783
RMW_t	0.014	0.000	0.531	-2.970	4.570	0.239	4.918
CMA_t	0.010	0.000	0.455	-5.310	2.480	-0.276	6.226
Panel B. Cross-section of 40 stocks: short sample (2015-2025)							
y_t	1.339	1.177	0.726	0.376	10.298	4.425	37.901
$\Delta \log(Volume_t)$	0.000	-0.015	0.353	-1.545	2.062	0.328	2.288
$\Delta \log(trades_t)$	0.000	-0.016	0.281	-1.234	1.771	0.509	2.799

Notes: Panel A reports summary statistics for the daily percentage realized volatility and set of predictors. See Appendix C for detailed definitions and sources of each series.The HAR variables are computed at forecast origin t using realized volatility through t. Panel B reports cross-sectional averages across the 40 individual stocks of ticker-level daily realized volatility and additional stock-specific time-series statistics over 2015–2025. Kurtosis is reported as excess kurtosis relative to normal distribution.

Figure 1: Sample autocorrelation functions for daily percentage realized volatility



Notes: The figure reports autocorrelation functions through lag 60 for daily percentage realized volatility, $100\sqrt{RV_t}$. The gray lines are the 40 individual stock autocorrelation functions. The green line is the average autocorrelation across the 40 stocks, the dashed blue and orange lines are the corresponding lag-specific minimum and maximum, and the red line is the S&P 500. Dashed horizontal lines denote approximate 95% white-noise bands, $\pm 1.96/\sqrt{T}$, computed using the smallest sample size among the plotted series.

the stock average and remains close to the upper envelope of the stock cross section over much of the reported lag range. This pattern supports the paper's modeling strategy: realized volatility displays strong persistence in all series, but the aggregate market index exhibits materially stronger long-lived dependence than the typical individual stock.

5 Empirical results

5.1 Results under the baseline

In the baseline specification, forecasts rely exclusively on the past history of realized volatility—summarized by the daily, weekly, and monthly components for the HAR-family models and by the corresponding univariate lag structure for ARFIMA.[7]

5.1.1 Out-of-sample performance results

Tables 2 and 3 provide a compact summary of the baseline results. Table 2 reports average out-of-sample MSFE, MAE, and QLIKE values by model, forecast horizon, and estimation window,

[7]Detailed results for both the baseline specification and specifications incorporating a EN expanded set of predictors are reported in Online Appendices A–E for S&P500 and Appendices H and I for cross-section and individual equities.

together with the corresponding MCS results. In this and analogous loss tables, underlined entries denote models included in the 5% MCS, while entries that are both boldfaced and underlined identify the lowest-loss model among the MCS members. Table 3 complements these average-loss comparisons by reporting the number of yearly test periods in which each model is the lowest-loss member of the 5% MCS. Two broad conclusions emerge. First, the substantive ranking of the models is remarkably stable across the expanding- and rolling-window estimation schemes. Second, the relative performance of the leading econometric models varies systematically with the forecast horizon: MSHAR performs best at short horizons, whereas ARFIMA generally leads at the monthly horizon. These conclusions are consistent across all three loss functions and therefore do not depend on a particular forecast-evaluation criterion.

Table 2: Average out-of-sample loss by window, horizon, and metric under the baseline information set

Window	Horizon	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel MSFE: average loss over 2006–2025											
	$h = 1$	0.093	0.092	0.091	**0.051**	0.104	0.102	0.093	0.103	0.096	0.096
Expanding	$h = 5$	0.167	0.166	0.167	**0.084**	0.104	0.183	0.164	0.178	0.179	0.177
	$h = 22$	0.246	0.257	0.252	0.155	**0.104**	0.276	0.261	0.270	0.275	0.276
	$h = 1$	0.093	0.092	0.092	**0.052**	0.103	0.118	0.093	0.109	0.105	0.105
Rolling 10y	$h = 5$	0.168	0.170	0.169	**0.096**	0.103	0.213	0.170	0.189	0.192	0.186
	$h = 22$	0.240	0.255	0.247	0.154	**0.104**	0.270	0.275	0.282	0.277	0.266
Panel MAE: average loss over 2006–2025											
	$h = 1$	0.179	0.177	0.178	**0.149**	0.198	0.183	0.180	0.179	0.177	0.177
Expanding	$h = 5$	0.241	0.239	0.241	**0.189**	0.198	0.247	0.239	0.238	0.240	0.241
	$h = 22$	0.300	0.305	0.301	0.240	**0.198**	0.313	0.303	0.299	0.298	0.307
	$h = 1$	0.178	0.176	0.177	**0.151**	0.194	0.188	0.177	0.181	0.180	0.180
Rolling 10y	$h = 5$	0.238	0.238	0.239	0.196	**0.194**	0.252	0.239	0.244	0.245	0.242
	$h = 22$	0.293	0.301	0.296	0.239	**0.195**	0.307	0.306	0.303	0.296	0.297
Panel QLIKE: average loss over 2006–2025											
	$h = 1$	0.586	0.581	0.585	**0.569**	0.591	0.582	0.583	0.581	0.581	0.580
Expanding	$h = 5$	0.623	0.617	0.623	**0.588**	0.591	0.621	0.617	0.619	0.620	0.619
	$h = 22$	0.664	0.666	0.666	0.615	**0.593**	0.675	0.672	0.678	0.675	0.673
	$h = 1$	0.586	0.580	0.586	**0.570**	0.590	0.584	0.581	0.582	0.581	0.581
Rolling 10y	$h = 5$	0.624	0.619	0.624	0.593	**0.590**	0.625	0.619	0.625	0.625	0.622
	$h = 22$	0.666	0.670	0.668	0.620	**0.592**	0.678	0.702	0.696	0.683	0.679

Notes: Entries are the average out-of-sample (2006-2025) losses under each loss measure. Underlined entries indicate models that belong to the 5% Model Confidence Set computed on the full pooled daily loss series for the corresponding window, horizon, and metric. Entries that are both boldfaced and underlined are the lowest-loss models among those in the MCS. Bootstrap design: stationary bootstrap, $B = 5000$, block length $\lfloor \sqrt{T} \rfloor$.

A useful starting point is the comparison across training-window designs. The identity of the top-ranked model is unchanged across expanding and rolling 10-year estimation in seven of the nine horizon–metric combinations reported in Table 2. The only departures arise at the intermediate horizon, $h = 5$, where the rolling window shifts the ranking modestly in favor of ARFIMA under MAE and QLIKE. Overall, however, the benchmark findings are qualitatively invariant to whether the models are estimated using the full historical sample available at each

forecast origin or only the most recent decade of data. This stability is important because it indicates that the main conclusions are not an artifact of a particular choice of training window.

At the one-day horizon, MSHAR is the clear benchmark leader. It delivers the lowest average loss in all six $h = 1$ cases in Table 2, covering the three loss functions under both estimation windows. The win-count evidence in Table 3 (based on yearly test results reported in detail in an Online Appendix) reinforces the same conclusion: MSHAR is the lowest-loss model within the MCS in 19 of 20 yearly test periods under expanding windows for each loss function, and in 18 of 20 yearly test periods under rolling windows. Hence, the superiority of MSHAR at the short horizon is not driven by a small number of crisis observations or unusual subsamples. Rather, it is a pervasive feature of the yearly out-of-sample record. This pattern is consistent with the view that short-run realized volatility is especially sensitive to regime changes and state-dependent nonlinearities.

An important qualification is that not all nonlinear econometric models perform equally well. Although THAR and STHAR also allow for regime dependence, they rarely improve materially on the linear HAR benchmark and almost never emerge as the lowest-loss member of the MCS. In the benchmark sample, THAR records no yearly wins, while STHAR records only a single win in the $h = 1$ panels. Their average losses are typically clustered near HAR rather than near MSHAR. A plausible interpretation is that the relevant nonlinearity in realized volatility is not well described by a single observed threshold or smooth transition variable. If short-horizon volatility is governed by recurrent latent states that alter both the level and persistence of volatility, then a Markov-switching specification has a natural advantage over deterministic threshold and smooth-transition rules tied to one observed driver. Put differently, the results suggest that allowing for nonlinearity *per se* is not sufficient; the form through which nonlinearity is introduced is empirically consequential.

The ranking reverses at the monthly horizon. For $h = 22$, ARFIMA records the lowest average loss in all six horizon–metric–window combinations and overwhelmingly dominates the yearly win counts, with 16 to 18 wins out of 20 depending on the loss function and training scheme. MSHAR remains the closest competitor at this horizon, but it is consistently second to ARFIMA. This result accords with the economic interpretation of volatility persistence: as the forecast horizon lengthens, long-memory dynamics become more important, while the incremental gains from modeling regime-switching behavior diminish relative to the gains from capturing highly persistent dependence in volatility.

The five-day horizon is the only case in which the benchmark ranking is not completely uniform. Under MSFE, MSHAR remains the best-performing model under both expanding and rolling estimation, both in terms of average loss and yearly win counts. Under expanding windows, MSHAR also retains the lowest average MAE and QLIKE, although the MAE win counts show that ARFIMA is nearly tied in practice. Under rolling 10-year estimation,

Table 3: Win counts: number of yearly test periods in which a model is the lowest-loss member of the MCS

Window	Horizon	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel MSFE: yearly win counts (out of 20)											
Expanding	$h = 1$	0	0	1	**19**	0	0	0	0	0	0
	$h = 5$	0	0	0	**12**	7	0	0	0	1	0
	$h = 22$	0	0	0	3	**16**	0	0	0	1	0
Rolling 10y	$h = 1$	0	0	1	**18**	0	0	0	0	1	0
	$h = 5$	0	0	0	**11**	8	0	0	0	0	1
	$h = 22$	0	0	0	3	**17**	0	0	0	0	0
Panel MAE: yearly win counts (out of 20)											
Expanding	$h = 1$	0	0	1	**19**	0	0	0	0	0	0
	$h = 5$	0	0	0	**8**	**8**	0	0	1	2	1
	$h = 22$	0	0	0	1	**17**	0	0	0	2	0
Rolling 10y	$h = 1$	0	0	1	**18**	0	0	0	0	1	0
	$h = 5$	0	0	0	5	**12**	0	0	0	1	2
	$h = 22$	0	0	0	1	**18**	0	0	0	1	0
Panel QLIKE: yearly win counts (out of 20)											
Expanding	$h = 1$	0	0	1	**19**	0	0	0	0	0	0
	$h = 5$	0	0	0	**11**	8	0	0	0	1	0
	$h = 22$	0	0	0	2	**17**	0	0	0	1	0
Rolling 10y	$h = 1$	0	0	1	**18**	0	0	0	0	1	0
	$h = 5$	0	0	0	8	**11**	0	0	0	0	1
	$h = 22$	0	0	0	2	**18**	0	0	0	0	0

Notes: For each year, window, horizon, and loss metric, a model receives a win if it is the boldfaced-and-underlined entry in the corresponding yearly table, i.e., the lowest-loss model among those included in the 5% Model Confidence Set. Yearly MSFE, MAE, QLIKE and MCS results are in Online Appendices. Each row sums to 20 yearly test periods (2006–2025). Boldfaced counts identify the largest win total within each row.

ARFIMA moves into first place under both MAE and QLIKE and also records the larger number of yearly wins for those criteria. The weekly horizon is therefore best viewed as a transition region: both regime dependence and long-memory persistence matter, and their relative importance depends somewhat on the loss function and the estimation window.

The baseline evidence also helps place the machine-learning models in perspective. The neural-network and boosting specifications occasionally register isolated yearly wins, but these are sparse and do not translate into systematic gains in average loss. More importantly, none of the machine-learning models consistently dominates THAR or STHAR, and none comes close to challenging MSHAR at $h = 1$ or ARFIMA at $h = 22$. In several panels, HAR, THAR and STHAR remain at least as competitive as the stronger machine-learning alternatives, even though they fail to outperform MSHAR or ARFIMA. The implication is that the benchmark ordering is not well described by a simple "econometric versus machine learning" contrast. Rather, predictive performance appears to depend on whether the model structure is aligned with the dominant feature of volatility at a given horizon: latent regime switching at short horizons and long-memory persistence at longer horizons.

Taken together, the results in Tables 2 and 3 establish a coherent benchmark for the remainder of the paper. The principal findings are not sensitive to the training-window design, they are robust across MSFE, MAE, and QLIKE, and they identify a simple horizon-specific pattern:

MSHAR is the preferred benchmark at short horizons, ARFIMA is the preferred benchmark at long horizons, and the weekly horizon lies between these two regimes. Equally important, the results distinguish between different forms of nonlinearity. Threshold and smooth-transition extensions of HAR provide only limited gains over the linear benchmark, whereas Markov-switching dynamics yield substantial improvements at short horizons. This distinction will be useful when interpreting the subsequent results for richer predictor sets and for the machine-learning specifications.

We report episode-specific results for four stress periods: the Global Financial Crisis (December 2007–June 2009), the European sovereign debt crisis (January 2010–June 2013), the COVID-19 shock (February 2020–April 2020), and the post-COVID inflation and monetary-tightening episode (June 2021–December 2024). These windows are intended to capture distinct volatility environments, ranging from systemic financial stress and sovereign-risk concerns to the abrupt pandemic shock and the subsequent period of elevated inflation, rising interest rates, and tighter monetary policy.

Table 4 reports episode-specific forecast losses under the rolling 10-year estimation window for four economically salient subperiods: the Great Financial Crisis, the European sovereign debt crisis, the COVID shock, and the recent inflationary episode. The corresponding episode table under the expanding-window design is reported in the Online Appendix. This presentation strikes a useful balance between transparency and economy. The rolling-window table is informative about performance under structural change and evolving local dynamics, whereas the expanding-window results are qualitatively similar and do not alter the main conclusions.

Three findings stand out. First, the episode-level evidence strongly reinforces the full-sample benchmark results reported earlier. At the one-day horizon, MSHAR is the best-performing model in every episode and under all three loss functions. At the 22-day horizon, ARFIMA is preferred in every episode and under all three loss functions. Hence, the short-horizon advantage of latent regime switching and the long-horizon advantage of long-memory dynamics are not artifacts of averaging across heterogeneous market environments. These ranking patterns remain visible even when attention is restricted to crisis and post-crisis subperiods.

Second, the five-day horizon again emerges as the main region of competition between model classes. During the Great Financial Crisis, MSHAR is preferred under MSFE, MAE, and QLIKE, and the same conclusion continues to hold during the European sovereign debt crisis under MSFE and MAE. By contrast, ARFIMA becomes the preferred specification at $h = 5$ during the COVID shock and the post-2021 inflationary episode under all three loss functions, while already edging ahead under QLIKE during the European sovereign debt crisis. This intermediate-horizon pattern is consistent with the benchmark results from the full sample: the weekly horizon appears to be the point at which short-run regime dependence and longer-

Table 4: Episode-specific forecast loss comparison under rolling 10Y window

Episode	Horizon	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel A: MSFE											
GFC	1	0.315	0.324	0.307	**0.139**	0.376	0.479	0.318	0.440	0.392	0.440
	5	0.511	0.538	0.514	**0.231**	0.376	0.782	0.520	0.683	0.719	0.715
	22	0.890	0.904	0.894	0.544	**0.374**	1.035	0.878	1.175	1.167	1.047
EDC	1	0.066	0.067	0.066	**0.044**	0.068	0.068	0.067	0.067	0.065	0.065
	5	0.112	0.113	0.112	**0.063**	0.068	0.117	0.113	0.115	0.116	0.118
	22	0.154	0.157	0.155	0.081	**0.068**	0.169	0.148	0.156	0.160	0.165
COVID	1	0.865	0.776	0.827	**0.343**	1.002	1.632	0.846	1.475	1.425	1.237
	5	2.169	2.129	2.161	1.368	**1.004**	2.726	2.245	2.833	2.553	2.392
	22	3.372	3.889	3.790	2.379	**0.783**	3.786	4.741	3.852	3.806	3.722
INF	1	0.059	0.054	0.058	**0.032**	0.056	0.060	0.055	0.053	0.053	0.053
	5	0.095	0.090	0.096	0.059	**0.056**	0.093	0.089	0.086	0.086	0.087
	22	0.131	0.120	0.131	0.098	**0.057**	0.123	0.124	0.124	0.117	0.119
Panel B: MAE											
GFC	1	0.307	0.306	0.306	**0.225**	0.333	0.370	0.309	0.350	0.336	0.349
	5	0.424	0.421	0.423	**0.266**	0.333	0.501	0.429	0.483	0.500	0.488
	22	0.546	0.589	0.546	0.382	**0.329**	0.590	0.540	0.650	0.648	0.602
EDC	1	0.165	0.166	0.166	**0.146**	0.178	0.167	0.165	0.165	0.164	0.165
	5	0.206	0.211	0.207	**0.175**	0.178	0.212	0.205	0.209	0.209	0.215
	22	0.261	0.260	0.263	0.204	**0.179**	0.271	0.247	0.254	0.259	0.280
COVID	1	0.629	0.575	0.619	**0.424**	0.641	0.798	0.608	0.766	0.746	0.692
	5	1.010	0.963	0.998	0.757	**0.648**	1.070	1.047	1.090	1.022	0.989
	22	1.259	1.407	1.328	1.027	**0.555**	1.337	1.737	1.293	1.286	1.212
INF	1	0.170	0.163	0.169	**0.139**	0.165	0.169	0.161	0.162	0.161	0.161
	5	0.221	0.213	0.222	0.186	**0.165**	0.217	0.209	0.206	0.207	0.209
	22	0.255	0.247	0.258	0.225	**0.166**	0.253	0.248	0.251	0.242	0.244
Panel C: QLIKE											
GFC	1	0.032	0.033	0.031	**0.016**	0.040	0.045	0.032	0.041	0.037	0.041
	5	0.064	0.066	0.064	**0.026**	0.040	0.093	0.065	0.083	0.085	0.084
	22	0.141	0.136	0.142	0.058	**0.039**	0.158	0.139	0.214	0.203	0.169
EDC	1	0.038	0.039	0.039	**0.032**	0.043	0.040	0.039	0.039	0.038	0.039
	5	0.063	0.064	0.064	0.044	**0.043**	0.065	0.063	0.066	0.067	0.066
	22	0.098	0.100	0.099	0.058	**0.044**	0.109	0.096	0.103	0.105	0.106
COVID	1	0.087	0.081	0.083	**0.036**	0.102	0.147	0.086	0.128	0.122	0.105
	5	0.342	0.359	0.348	0.113	**0.102**	0.438	0.351	0.485	0.424	0.382
	22	0.985	1.175	1.097	0.288	**0.067**	1.211	1.378	1.327	1.269	1.269
Inf	1	0.046	0.045	0.046	**0.033**	0.046	0.046	0.044	0.043	0.043	0.044
	5	0.076	0.073	0.076	0.057	**0.046**	0.075	0.074	0.071	0.071	0.073
	22	0.103	0.100	0.104	0.085	**0.045**	0.103	0.103	0.102	0.096	0.098

Notes: Panel A reports MSFE, Panel B reports MAE, and Panel C reports QLIKE. Within each panel, rows are organized by selected episode and forecast horizon $h = 1$, $h = 5$, and $h = 22$. Entries are computed from the daily forecast files using only observations whose dates fall within the stated episode. Underlined entries indicate models included in the 5% Model Confidence Set for the corresponding episode, horizon, and metric, using aligned daily losses. Entries that are both underlined and boldfaced are the lowest-loss models in that row among the models included in the MCS. Bootstrap design: stationary bootstrap, $B = 5000$, block length $\lfloor \sqrt{T} \rfloor$.

memory persistence are most closely balanced, so that their relative importance becomes more sensitive to the underlying volatility environment.

Third, the episode results are informative not only about model ranking but also about the overall difficulty of the forecasting problem. The level of the loss varies substantially across subperiods, even when the identity of the best-performing model does not. In particular, losses are markedly larger during the Great Financial Crisis and, especially, during the COVID shock than during the European sovereign debt crisis or the inflationary episode. This feature is visible across MSFE, MAE, and QLIKE. For example, under MSFE at $h = 22$, the best loss rises from 0.057 in the inflationary episode and 0.068 during the European sovereign debt crisis to 0.374 during the Great Financial Crisis and 0.783 during the COVID shock. Thus, the episode analysis clarifies an important distinction: model ranking and forecast difficulty should be interpreted separately. Some environments are simply much harder to forecast than others.

The episode table also sharpens the interpretation of the nonlinear and machine-learning results. Although THAR and STHAR allow for regime dependence, they do not emerge as the leading specification in any of the reported episode–horizon–loss combinations. Their losses are often close to those of HAR and, in some rows, modestly below the linear benchmark, but they do not systematically challenge MSHAR at short horizons or ARFIMA at long horizons. A plausible interpretation is that the relevant nonlinearity in realized volatility is not well captured by deterministic threshold or smooth-transition mechanisms alone. Instead, the data appear to favor a more flexible latent-state representation at short horizons, as embodied by the Markov-switching HAR specification. The machine-learning models deliver a similar message. They occasionally narrow the gap relative to HAR and sometimes outperform THAR or STHAR, but they do not dominate the benchmark comparisons in any systematic way. The evidence therefore does not support a general claim that either nonlinear extensions of HAR or machine-learning methods *per se* improve forecast performance. Rather, the form of the dynamics matters: latent regime switching is most valuable at short horizons, whereas long-memory persistence is most valuable at longer horizons.

Overall, the episode-level evidence strengthens the benchmark conclusions while adding an important nuance. The main ranking is remarkably stable across very different market environments, but the absolute size of the forecasting errors varies sharply across episodes, and the five-day horizon remains the principal region in which the balance between regime-switching dynamics and persistent dependence becomes most sensitive to the underlying state of the volatility process.

5.1.2 Diebold–Mariano test results

Table 5 summarizes the full-sample two-sided Diebold–Mariano (DM) tests under the baseline HAR predictor set. To keep the main text concise, the table reports only the pairwise com-

parisons that are most relevant for the benchmark narrative. The complete full-sample DM matrices by loss function and training window are reported in the Online Appendix.

The DM evidence strongly reinforces the conclusions from the loss and Model Confidence Set tables. Panel A shows that the short-horizon advantage of MSHAR is not a marginal feature of one particular loss criterion. MSHAR significantly outperforms HAR, THAR, and STHAR under all three loss functions at all three horizons under both the expanding and rolling 10-year estimation windows. The statistical evidence therefore points to a sharp distinction between latent Markov switching and the deterministic threshold and smooth-transition mechanisms embedded in THAR and STHAR. The data do not merely reward nonlinearity *per se*; they reward a particular form of nonlinearity. Panel A also clarifies the horizon-specific role of ARFIMA. Relative to HAR, ARFIMA is significantly worse at $h = 1$ under both training schemes and all three loss functions, but significantly better at $h = 5$ and $h = 22$ in every case. When ARFIMA is compared directly with MSHAR, the horizon pattern becomes even sharper: MSHAR dominates ARFIMA at $h = 1$ under all losses and both windows, whereas ARFIMA dominates MSHAR at $h = 22$ in all cases. The only horizon at which the two leading econometric models are not cleanly separated is $h = 5$, where MSHAR wins two of the three loss functions under the expanding window, ARFIMA wins only QLIKE under the rolling window, and the remaining comparisons are not statistically distinguishable.

Panel B sharpens the comparison with machine-learning models from the perspective of the strongest short-horizon econometric benchmark. Across the full sample, MSHAR significantly outperforms MLP, LSTM, LSTM-A, and GRU-A in all 18 horizon–loss–window comparisons, and it significantly outperforms XGB in 17 of the 18 comparisons, with the sole exception being QLIKE at $h = 5$ under the rolling 10-year window. Thus, the machine-learning models do not merely fail to rank first on average; they also generally fail to eliminate the statistically meaningful advantage of the best-performing short-horizon econometric model.

Panel C adds a complementary perspective by comparing ARFIMA with the machine-learning models. Here the pattern is notably more horizon-dependent. At the one-day horizon, ARFIMA is often significantly worse than several machine-learning models, especially MLP, LSTM-A, and GRU-A, and its comparison with XGB is mixed across windows and loss functions. By contrast, once the horizon extends to five or 22 trading days, ARFIMA significantly outperforms every machine-learning specification under both training windows and all three loss functions. Hence, the statistical evidence against the machine-learning models is even broader than the ranking tables alone suggest: MSHAR dominates them where short-run regime changes matter most, whereas ARFIMA dominates them where longer-memory dynamics become most relevant.

Taken together, the DM results provide a useful statistical counterpart to the loss and MCS evidence. Under the baseline HAR predictor set, the short-horizon gains are decisively

concentrated in MSHAR, the long-horizon gains are decisively concentrated in ARFIMA, and the five-day horizon remains the main intermediate case in which the two leading econometric models are not uniformly separable across all criteria. This is precisely the pattern one would expect if latent regime changes matter most for very short-run volatility forecasting, whereas long-memory dynamics become increasingly important as the forecast horizon lengthens.

Table 5: Compact summary of full-sample two-sided Diebold–Mariano tests under the baseline information set.

Pairwise comparison	Window	$h = 1$	$h = 5$	$h = 22$
Panel A: Main econometric benchmark comparisons				
HAR vs MSHAR	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
THAR vs MSHAR	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
STHAR vs MSHAR	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
HAR vs ARFIMA	Expanding	HAR (3/3)	ARFIMA (3/3)	ARFIMA (3/3)
	Rolling 10-year	HAR (3/3)	ARFIMA (3/3)	ARFIMA (3/3)
MSHAR vs ARFIMA	Expanding	MSHAR (3/3)	MSHAR (2/3)	ARFIMA (3/3)
	Rolling 10-year	MSHAR (3/3)	ARFIMA (1/3)	ARFIMA (3/3)
Panel B: MSHAR versus machine-learning models				
MSHAR vs XGB	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (2/3)	MSHAR (3/3)
MSHAR vs MLP	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
MSHAR vs LSTM	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
MSHAR vs LSTM-A	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
MSHAR vs GRU-A	Expanding	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
	Rolling 10-year	MSHAR (3/3)	MSHAR (3/3)	MSHAR (3/3)
Panel C: ARFIMA versus machine-learning models				
ARFIMA vs XGB	Expanding	XGB (2/3)	ARFIMA (3/3)	ARFIMA (3/3)
	Rolling 10-year	ARFIMA (1/3), XGB (2/3)	ARFIMA (2/3)	ARFIMA (3/3)
ARFIMA vs MLP	Expanding	MLP (3/3)	ARFIMA (3/3)	ARFIMA (3/3)
	Rolling 10-year	MLP (3/3)	ARFIMA (3/3)	ARFIMA (3/3)
ARFIMA vs LSTM	Expanding	LSTM (2/3)	ARFIMA (3/3)	ARFIMA (3/3)
	Rolling 10-year	LSTM (2/3)	ARFIMA (3/3)	ARFIMA (3/3)
ARFIMA vs LSTM-A	Expanding	LSTM-A (3/3)	ARFIMA (3/3)	ARFIMA (3/3)
	Rolling 10-year	LSTM-A (2/3)	ARFIMA (3/3)	ARFIMA (3/3)
ARFIMA vs GRU-A	Expanding	GRU-A (3/3)	ARFIMA (3/3)	ARFIMA (3/3)
	Rolling 10-year	GRU-A (2/3)	ARFIMA (3/3)	ARFIMA (3/3)

Notes: Each entry reports which model significantly outperforms the other model in the stated pairwise comparison, together with the number of loss functions for which the two-sided Diebold–Mariano null of equal predictive accuracy is rejected at the 5% level. Counts are out of three loss functions (MSFE, MAE, and QLIKE). For example, "MSHAR (2/3)" means that MSHAR significantly outperforms the comparator under two of the three loss functions, while the remaining loss function yields no statistically significant difference. An entry such as "ARFIMA (1/3), XGB (2/3)" means that the direction of statistically significant outperformance differs across loss functions. The underlying full-sample DM matrices by loss function and training window are reported in the Online Appendix.

5.1.3 Realized utility

Table 6 shows that the realized-utility evidence closely mirrors the loss-based and Diebold–Mariano results. At the one-day horizon, MSHAR delivers the highest utility under both training schemes, with RU values of 3.931 under the expanding window and 3.929 under the rolling 10-year window. Relative to HAR, these gains are modest in absolute terms—about 0.024 and 0.022 utility points, respectively—but they are remarkably consistent and again indicate that short-run volatility forecasts benefit from latent regime switching.

At the intermediate horizon $h = 5$, the ranking remains transitional. Under the expanding window, MSHAR continues to generate the highest utility, with RU = 3.896 versus 3.890 for ARFIMA and 3.831 for HAR. Under the rolling 10-year window, however, ARFIMA moves into first place with RU = 3.891, narrowly ahead of MSHAR at 3.884 and clearly above HAR at 3.828. This pattern is fully consistent with the baseline loss results: the five-day horizon is the one region in which short-run regime dependence and longer-memory persistence are both empirically relevant.

At the monthly horizon $h = 22$, ARFIMA is the clear economic winner under both training windows. Its full-sample RU equals 3.890 under both expanding and rolling estimation, compared with 3.846 and 3.832 for MSHAR and only 3.739 and 3.730 for HAR. The implied utility gains relative to HAR—roughly 0.151 under expanding and 0.160 under rolling estimation—are economically more pronounced than at shorter horizons and reinforce the conclusion that long-memory dynamics become increasingly important as the forecast horizon lengthens.

The threshold and smooth-transition HAR models yield utility levels that are generally very close to HAR, and they do not deliver economically meaningful improvements over the linear benchmark. This echoes the earlier loss-based evidence that allowing for nonlinearity *per se* is not sufficient; rather, the form through which nonlinearity enters the model matters empirically. In these data, latent Markov switching is considerably more valuable than deterministic threshold or smooth-transition specifications.

The machine-learning models likewise fail to overturn the benchmark ranking. At $h = 1$, the best-performing ML models are close to HAR but remain below MSHAR under both training windows. At $h = 5$, the ML models remain below the leading econometric specifications and typically do not improve even on HAR by much. At $h = 22$, the gap widens markedly: all ML models lie well below both MSHAR and ARFIMA, and several recurrent architectures fall noticeably short of even the linear HAR benchmark under the rolling-window design. Accordingly, the RU evidence supports the same central conclusion as the statistical loss measures: under the baseline HAR predictor set, economic gains come primarily from MSHAR at short horizons and from ARFIMA at long horizons, while threshold-based nonlinear models and ML alternatives do not generate systematic improvements over the strongest econometric benchmarks.

As with the loss metrics, the detailed yearly and episode-specific RU tables are presented in the Online Appendix. Results in Appendix document how the level of realized utility varies across subperiods—especially during crisis episodes such as COVID—rather than to alter the core ranking, which is already summarized transparently by the full-sample results in Table 6.

Table 6: Full-sample realized utility under baseline information set

Model	$h = 1$	$h = 5$	$h = 22$
Panel A. Expanding window			
HAR	3.907	3.831	3.739
THAR	3.906	3.832	3.726
STHAR	3.908	3.831	3.735
MSHAR	**3.931**	**3.896**	3.846
ARFIMA	3.890	3.890	**3.890**
XGB	3.904	3.824	3.707
MLP	3.902	3.830	3.715
LSTM	3.903	3.824	3.691
LSTM-A	3.905	3.822	3.696
GRU-A	3.906	3.826	3.705
Panel B. Rolling 10-year window			
HAR	3.907	3.828	3.730
THAR	3.906	3.826	3.715
STHAR	3.907	3.826	3.725
MSHAR	**3.929**	3.884	3.832
ARFIMA	3.891	**3.891**	**3.890**
XGB	3.900	3.815	3.695
MLP	3.906	3.826	3.633
LSTM	3.903	3.809	3.645
LSTM-A	3.904	3.808	3.674
GRU-A	3.905	3.816	3.687

Notes: Entries report full-sample realized utility (RU) computed over the 2006–2025 out-of-sample period under the baseline HAR predictor set. The daily utility measure is defined as $RU_t = 100\left(0.08\sqrt{RV_t/\widehat{RV}_t} - 0.04(RV_t/\widehat{RV}_t)\right)$, so that a perfect volatility forecast yields RU= 4.000. Boldface identifies the highest RU within each horizon and training-window panel. Detailed yearly and episode-specific RU tables are reported in the Online Appendix.

5.1.4 Value-at-Risk exceedances

Table 7 reports full-sample 99% VaR exceedance rates obtained from the filtered historical simulation procedure under the baseline HAR predictor set. In contrast to the loss-based and utility-based exercises, the VaR results are primarily about *tail calibration* rather than broad ranking across all models. The key message is that VaR performance depends strongly on both the forecast horizon and the estimation window. In particular, the five-day horizon is the most

stable from a coverage perspective, whereas the one-day and 22-day horizons display larger departures from the nominal 1% exceedance rate under the rolling 10-year window.

Under the expanding window, the one-day results are relatively well behaved. MSHAR delivers an exceedance rate of 0.99%, essentially on target, and is the only model at $h = 1$ for which neither the unconditional- nor conditional-coverage test indicates a statistically significant departure from nominal coverage. ARFIMA also performs reasonably well at 1.16%, with no significant coverage-test departure, whereas HAR, THAR, STHAR, and most machine-learning models exhibit exceedance rates somewhat above the nominal level and are more often flagged by the Christoffersen conditional-coverage test. At the five-day horizon, however, the differences across models narrow substantially: all exceedance rates lie in a relatively tight range around 1.23–1.42%, and the full-sample coverage tests do not indicate significant departures for any model. The 22-day expanding-window results remain above the nominal 1% rate, but the full-sample backtests again do not indicate statistically significant departures from the coverage nulls. Taken together, the expanding-window evidence suggests that the baseline VaR implementation is reasonably calibrated in the full sample, with the strongest one-day calibration delivered by MSHAR.

The rolling 10-year results show larger calibration departures at the short and long horizons. At $h = 1$, all models cluster between 1.35% and 1.48%, so exceedance rates remain low in absolute terms but are above the nominal 1% benchmark; the unconditional- and conditional-coverage tests indicate statistically significant departures for all models in the full sample. At $h = 22$, exceedance rates rise further, ranging from 1.70% for MLP to 2.45% for STHAR, and the coverage tests flag departures from nominal calibration for most specifications. By contrast, the five-day horizon again appears considerably more stable: exceedance rates remain in a narrow range around 1.33–1.46%, and the full-sample tests do not indicate significant coverage departures for any rolling-window model. Thus, the most robust feature of the VaR evidence is not a single model ranking, but rather the horizon-specific calibration pattern whereby the intermediate horizon is comparatively well behaved while the rolling-window implementation exhibits larger deviations from nominal coverage at the shortest and longest horizons.

The model ranking under VaR calibration is therefore related to, but not identical with, the ranking obtained from forecast-loss measures and realized utility. MSHAR remains especially notable at the one-day horizon under the expanding window, where it achieves essentially exact nominal coverage in the full sample. At the 22-day horizon, ARFIMA yields the lowest exceedance rate under both windows, although its rolling-window exceedance rate remains above the 1% target and is flagged by the unconditional-coverage test. Threshold and smooth-transition HAR models do not improve materially on HAR itself from a VaR-calibration perspective, while the machine-learning models do not provide systematic tail-risk gains relative to the strongest econometric benchmarks.

The year-by-year results, reported in the Online Appendix, further show that coverage-test departures are not evenly distributed over time. Rather, statistically significant departures from unconditional and conditional coverage occur more frequently during elevated-volatility periods, especially in 2007–2008 and 2020, with additional pockets visible in later stressed conditions such as 2025 for some specifications. In contrast, during calmer market environments—such as 2012–2014 and 2017—exceedance rates are often close to 0.00 for several models, suggesting that VaR forecasts can become conservative when tail risk is muted. This temporal asymmetry is economically intuitive and helps explain why full-sample coverage-test flags are more pronounced for estimation schemes—particularly the rolling 10-year window—that may be less able to adjust smoothly to abrupt changes in return tails and, at the same time, may overstate tail risk during tranquil periods.

Table 7: Full-sample VaR exceedance rates under filtered historical simulation at the 99% level

Horizon	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel A. Expanding window										
$h = 1$	1.28†	1.27†	1.28†	0.99	1.16	1.16†	1.20†	1.23†	1.27†	1.27†
$h = 5$	1.42	1.33	1.39	1.37	1.31	1.29	1.40	1.23	1.29	1.27
$h = 22$	1.75	1.68	1.77	1.74	1.57	1.53	1.21	1.51	1.55	1.60
Panel B. Rolling 10-year window										
$h = 1$	1.48*†	1.44*†	1.46*†	1.35*†	1.48*†	1.37*†	1.39*†	1.42*†	1.42*†	1.44*†
$h = 5$	1.46	1.40	1.46	1.42	1.42	1.33	1.40	1.35	1.40	1.33
$h = 22$	2.41*†	2.38*†	2.45*†	2.32*	1.85*	2.15*†	1.70*†	1.87*†	1.79*†	2.08*†

Notes: Entries report full-sample VaR exceedance rates in percent using filtered historical simulation (FHS) at tail probability $\alpha = 0.01$. The nominal exceedance rate is therefore 1%. For each model m and horizon h, the historical innovation pool is model-specific and uses standardized residuals based on that model's own volatility forecast; historical information is truncated at $d - h$ to avoid look-ahead bias. For $h = 5$ and $h = 22$, the Kupiec unconditional-coverage (UC) and Christoffersen conditional-coverage (CC) tests are computed on non-overlapping target dates. An asterisk (*) denotes rejection of the Kupiec null at the 5% level, and a dagger (†) denotes rejection of the Christoffersen null at the 5% level. Detailed year-by-year results for both estimation windows are reported in the Online Appendix.

5.1.5 Diebold–Mariano tests under the FZ0 joint VaR–ES loss

Table 8 reports the full-sample Diebold–Mariano (DM) statistics under the rolling 10-year estimation window using the FZ0 joint VaR–ES loss. We focus on this specification in the main text because the rolling window is the more demanding design under potential structural change and, under the FZ0 loss, it also provides the clearest tail-risk counterpart to the main forecasting results. The corresponding expanding-window full-sample table, together with the episode-specific DM results for both window schemes, is reported in the Online Appendix. Overall, the FZ0-based DM evidence is consistent with the broader benchmark ranking established earlier, although the statistical separation across models is naturally weaker here than under the volatility-loss criteria.

At the one-day horizon, the rolling-window FZ0 results strongly favor MSHAR. In particular, MSHAR significantly outperforms HAR, THAR, STHAR, and ARFIMA, and it also significantly dominates all of the machine-learning benchmarks reported in the table. This finding is important because it shows that the short-horizon advantage of MSHAR extends beyond conventional forecast-loss criteria and remains visible when evaluation is based directly on joint tail-risk forecasts for VaR and expected shortfall. At the same horizon, ARFIMA performs poorly relative to the leading benchmark models: HAR, THAR, STHAR, and MSHAR all dominate ARFIMA, with several differences statistically significant. Thus, under the FZ0 criterion the short-horizon evidence is unambiguous in identifying MSHAR as the strongest benchmark specification.

At the five-day horizon, the ranking becomes less sharply separated. MSHAR and ARFIMA remain competitive relative to HAR-type alternatives and the machine-learning models, but the pairwise differences among the leading econometric models are generally no longer statistically significant under the rolling window. This intermediate-horizon result is consistent with the earlier loss-based evidence, which also suggested that the five-day horizon is the main zone where short-run regime dependence and longer-memory dynamics are both relevant. The expanding-window DM results reported in the Online Appendix reinforce the same qualitative message, although they deliver somewhat sharper evidence in favor of MSHAR at this horizon, including a statistically significant advantage of MSHAR over ARFIMA in the full sample.

At the 22-day horizon, the full-sample rolling-window FZ0 table does not deliver a statistically decisive ranking among HAR, MSHAR, and ARFIMA. Nevertheless, the pattern of DM statistics is still informative. First, the short-horizon weakness of ARFIMA does not carry over to the long horizon. Second, ARFIMA tends to compare more favorably against the machine-learning models at $h = 22$, and in several cases the ML alternatives are significantly inferior. Taken together with the volatility-loss results, this suggests that ARFIMA remains the most credible long-horizon benchmark even when the joint quality of VaR and ES forecasts is used as the evaluation criterion, although the full-sample FZ0-based pairwise evidence is less sharp than at $h = 1$.

The episode-specific DM results in the Online Appendix add useful nuance to the full-sample picture. In particular, the strongest longer-horizon evidence in favor of ARFIMA emerges in the European sovereign debt episode, where under the rolling window ARFIMA significantly outperforms HAR, THAR, STHAR, and MSHAR at $h = 22$. By contrast, the GFC and COVID episode tables exhibit much weaker statistical separation, especially at the medium and long horizons, which is unsurprising given the shorter effective samples and the greater difficulty of tail-risk evaluation in abrupt crisis environments. The inflation episode also points to improved relative performance of ARFIMA at the long horizon, although there the strongest statistically significant differences are concentrated in comparisons with the machine-learning

Table 8: Diebold–Mariano tests for FZ0 VaR–ES loss, full sample under rolling10y window

Panel A: $h = 1$

	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
HAR		1.165	1.427	-2.188*	9.233*	3.338*	4.145*	4.071*	4.196*	3.889*
THAR			0.574	-2.355*	8.492*	3.176*	3.759*	3.716*	3.886*	3.553*
STHAR				-2.628*	7.012*	3.040*	3.421*	3.573*	3.868*	3.408*
MSHAR					5.251*	4.192*	3.813*	4.435*	4.769*	4.472*
ARFIMA						0.273	-2.252*	-2.173*	-1.429	-1.654
XGB							-1.381	-1.918	-1.577	-1.970*
MLP								0.094	0.622	0.473
LSTM									2.657*	1.165
LSTM-A										-0.503

Panel B: $h = 5$

	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
HAR		3.076*	-1.084	-0.146	0.855	2.775*	2.572*	3.374*	4.280*	3.483*
THAR			-3.338*	-1.119	-0.341	2.507*	1.446	2.994*	3.233*	2.779*
STHAR				-0.050	0.939	2.830*	2.658*	3.411*	4.344*	3.569*
MSHAR					1.336	2.255*	1.519	2.852*	3.317*	2.541*
ARFIMA						1.946	1.015	2.399*	2.859*	2.096*
XGB							-2.059*	0.197	-0.542	-1.412
MLP								2.107*	2.454*	1.868
LSTM									-0.538	-1.411
LSTM-A										-1.217

Panel C: $h = 22$

	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
HAR		0.188	1.249	-0.854	-0.246	1.793	-1.352	2.258*	2.059*	1.944
THAR			0.056	-0.730	-0.244	1.861	-1.563	2.246*	2.050*	1.977*
STHAR				-0.928	-0.321	1.782	-1.379	2.255*	2.054*	1.933
MSHAR					0.679	1.586	-1.010	2.043*	1.903	1.676
ARFIMA						1.430	-1.041	1.968*	1.845	1.513
XGB							-2.089*	1.669	1.727	-0.845
MLP								2.175*	2.053*	2.035*
LSTM									0.914	-2.292*
LSTM-A										-2.030*

Notes: Entries report the Diebold–Mariano statistic for pairwise comparisons under the FZ0 joint VaR–ES loss. Only the upper-triangular elements are shown. A starred entry indicates rejection at the 5% level for the two-sided Diebold–Mariano test. Negative (positive) values favor the column (row) model. For $h = 5$ and $h = 22$, the DM test uses the horizon argument h, the Bartlett long-run variance estimator, and the Harvey correction. VaR and ES forecasts are computed using filtered historical simulation, where the historical innovation pool for each model and horizon is formed from the model's own standardised residuals $\hat{z}_t^{(h)} = r_t^{(h)}/\hat{\sigma}_{t,h}^{(m)}$, with $\hat{\sigma}_{t,h}^{(m)} = \sqrt{\widehat{RV}_{t,h}^{(m)}}/100$, and the historical window truncated at $d - h$ to avoid look-ahead.

models rather than with the leading econometric alternatives. Across episodes, however, one conclusion remains stable: the one-day horizon consistently favors MSHAR, whereas the longer-horizon results are more supportive of ARFIMA, particularly in episodes where persistent tail dynamics appear more important.

Taken together, the FZ0-based DM tests provide a tail-risk-based confirmation of the paper's main horizon-dependent message. MSHAR remains the clearest short-horizon benchmark, while ARFIMA becomes progressively more competitive as the forecast horizon lengthens and, in some episodes, emerges as the dominant long-horizon model. The expanding-window and episode results reported in the Online Appendix are qualitatively consistent with this interpretation, although the exact pattern of statistically significant pairwise rejections varies across windows and subperiods.

5.2 Results under extended information set

Table 9 summarizes the full-sample results under the screened additional-predictor set. Two conclusions emerge immediately. First, the extending the set of predictors leaves the main horizon-dependent ranking largely intact. At the one-day horizon, MSHAR remains the clear leader under both estimation windows and all three loss functions. At the 22-day horizon, ARFIMA remains the clear leader under both windows and all three loss functions. Second, the five-day horizon again behaves as the transition case. Under MSFE, MSHAR remains the lowest-loss model under both windows, whereas under MAE and QLIKE ARFIMA moves into first place under both windows. Thus, the additional predictors do not overturn the core benchmark message; rather, they sharpen the intermediate-horizon split between regime-switching dynamics and long-memory persistence.

Relative to the HAR-only baseline summarized in Table 2, the model ordering is remarkably stable. The identity of the best model is unchanged in 16 of the 18 window–horizon–metric cells. The only changes occur under the expanding window at $h = 5$, where ARFIMA overtakes MSHAR under MAE and QLIKE. In all other cases, the leading model is the same as under the baseline specification: MSHAR at $h = 1$, MSHAR at $h = 5$ under MSFE, and ARFIMA at $h = 22$. This stability indicates that the gains from adding the screened extended predictors are modest relative to the gains obtained from capturing the correct dynamic structure of volatility itself.

A further point worth emphasizing is that the extending the set of predictors does not, in general, lower forecast losses relative to the baseline. Comparing the minimum loss attained in each cell of Table 9 with its counterpart in Table 2, the extending the set of predictors beyond past history of realzied volatility delivers weakly *higher* minimum losses in all nine expanding-window cells. The deterioration is especially visible at the long horizon, where the best QLIKE value rises from 0.593 under the HAR benchmark to 0.629 under the extended-information set.

Table 9: Average out-of-sample loss by window, horizon, and metric under extended set of predictors

Window	Horizon	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel MSFE: average loss over 2006–2025											
Expanding	$h = 1$	0.086	0.087	0.086	**0.058**	0.132	0.094	0.127	0.086	0.091	0.088
Expanding	$h = 5$	0.178	0.189	0.188	**0.095**	0.104	0.186	0.298	0.172	0.176	0.172
Expanding	$h = 22$	0.384	0.436	0.443	0.217	**0.127**	0.292	0.934	0.276	0.305	0.270
Rolling 10y	$h = 1$	0.086	0.085	0.086	**0.053**	0.156	0.104	0.082	0.100	0.100	0.093
Rolling 10y	$h = 5$	0.166	0.173	0.174	**0.099**	0.103	0.210	0.176	0.193	0.194	0.188
Rolling 10y	$h = 22$	0.254	0.265	0.264	0.155	**0.104**	0.295	0.268	0.272	0.277	0.272
Panel MAE: average loss over 2006–2025											
Expanding	$h = 1$	0.173	0.176	0.175	**0.151**	0.230	0.179	0.202	0.175	0.179	0.175
Expanding	$h = 5$	0.245	0.250	0.250	0.195	**0.192**	0.254	0.280	0.240	0.245	0.238
Expanding	$h = 22$	0.342	0.351	0.357	0.256	**0.207**	0.329	0.389	0.298	0.314	0.299
Rolling 10y	$h = 1$	0.172	0.174	0.174	**0.151**	0.262	0.180	0.181	0.182	0.183	0.177
Rolling 10y	$h = 5$	0.240	0.245	0.244	0.201	**0.193**	0.259	0.255	0.249	0.250	0.248
Rolling 10y	$h = 22$	0.303	0.314	0.313	0.241	**0.191**	0.321	0.311	0.299	0.303	0.303
Panel QLIKE: average loss over 2006–2025											
Expanding	$h = 1$	0.583	0.584	0.584	**0.575**	0.674	0.585	0.591	0.583	0.584	0.584
Expanding	$h = 5$	0.625	0.632	0.633	0.594	**0.593**	0.628	0.630	0.623	0.625	0.623
Expanding	$h = 22$	0.854	0.805	0.809	0.663	**0.629**	0.675	0.683	0.673	0.676	0.692
Rolling 10y	$h = 1$	0.583	0.584	0.585	**0.575**	0.760	0.585	0.586	0.585	0.586	0.584
Rolling 10y	$h = 5$	0.623	0.627	0.627	0.601	**0.596**	0.633	0.629	0.630	0.631	0.629
Rolling 10y	$h = 22$	0.674	0.678	0.678	0.624	**0.593**	0.685	0.699	0.698	0.711	0.699

Notes: Entries are the average out-of-sample (2006-2025) losses for each model under extended set of predictors. Underlined entries indicate models that belong to the 5% Model Confidence Set computed on the full pooled daily loss series for the corresponding window, horizon, and metric. Entries that are both boldfaced and underlined are the lowest-loss models among those in the MCS.

Under rolling 10-year estimation with extended predictors, HAR produces only very limited improvements: MAE falls slightly at $h = 5$ and $h = 22$, and MSFE at $h = 22$ is essentially unchanged, while the remaining cells again display slightly higher losses. Hence, the additional predictors do not generate a broad-based improvement in forecasting accuracy and may, in some cases, add estimation noise rather than economically useful signal.

The relative performance of the nonlinear deterministic threshold models also changes little under the extended predictor set. THAR and STHAR remain very close to HAR across windows, horizons, and metrics, and they do not challenge either MSHAR at short horizons or ARFIMA at long horizons. The machine-learning models likewise fail to deliver systematic gains. A few isolated cells show ML models matching or narrowly beating the leading econometric specification—for example, LSTM is marginally best during the COVID episode at $h = 1$ under the expanding window, and some recurrent architectures remain competitive in a handful of yearly rows—but these exceptions do not alter the overall ranking. The broad lesson is therefore the same as under the benchmark specification: nonlinearity *per se* is not sufficient, and the form through which nonlinearity enters the model remains empirically decisive.

The episode-specific tables, although not reported in the paper for reasons of space, broadly reinforce the full-sample interpretation and are available upon request. In the GFC, MSHAR dominates at $h = 1$ and $h = 5$, whereas ARFIMA dominates at $h = 22$ under both windows and

all three loss functions. The European sovereign debt episode delivers the same clean short- versus long-horizon split, while the five-day horizon is mixed between MSHAR and ARFIMA depending on the loss function and window. The COVID episode is the main exception: under the expanding window, LSTM is preferred at $h = 1$, whereas ARFIMA clearly dominates at $h = 5$ and $h = 22$ under both window schemes. Finally, in the inflationary episode, MSHAR again dominates at $h = 1$, ARFIMA dominates at $h = 22$, and the five-day horizon is mostly favorable to MSHAR, with only minor exceptions under expanding-window QLIKE. These episode patterns therefore reinforce the view that the main message is horizon-dependent rather than episode-specific.

5.3 Do the main findings generalize to individual equities?

The preceding analysis establishes the relative forecasting performance of the models for aggregate market volatility. An important question is whether the resulting conclusions reflect features specific to index-level realized volatility or extend more broadly to the volatility dynamics of individual equities. We examine this question using the full set of 40 individual U.S. equities available in the VOLARE stock file. This cross-sectional analysis provides a distinct setting in which to assess the generality of the index-level findings, given the greater importance of firm-specific information, idiosyncratic shocks, and heterogeneity in volatility dynamics at the stock level. The S&P 500 remains the market-index series analyzed in the preceding sections and is excluded from the cross-sectional aggregates reported here. Accordingly, the win shares, MCS inclusion frequencies, loss ratios, and cross-sectional DM summaries in this section are calculated exclusively across the 40 individual equities.

The stock-level design differs from the long-sample S&P 500 baseline in two respects. First, the VOLARE stock histories begin in 2015, whereas the main S&P 500 analysis uses a substantially longer sample from 1996 to 2025. Second, because the individual-stock histories are shorter, the rolling-window stock exercises use a 5-year estimation window rather than the 10-year rolling window used in the long-sample S&P 500 analysis. The expanding-window and rolling-window stock results should therefore be interpreted as evidence from a common 2015–2025 individual-equity sample, not as direct replacements for the long historical market-index results.

To make this distinction transparent, the Online Appendix reports a matched-sample S&P 500 analysis alongside the ticker-by-ticker stock results. Specifically, the appendix tables include the S&P 500 under the same 2015–2025 design used for the individual equities, including the 5-year rolling-window specification. These matched-sample S&P 500 results are not used when computing the 40-stock cross-sectional summaries in the main text. Instead, they serve a separate diagnostic purpose: they show whether the conclusions from the long-sample S&P 500 exercise are sensitive to shortening the sample and changing the rolling-window length to match

the stock-level design. The paper therefore contains two complementary S&P 500 comparisons: the main long-sample S&P 500 evidence based on 1996–2025, and the Online Appendix matched-sample S&P 500 evidence based on the 2015–2025 stock-sample design.

The expanded stock-level exercise keeps the model comparison as close as possible to the baseline forecasting analysis. For each stock, we estimate the same econometric and machine-learning model classes and evaluate forecasts at horizons $h = 1$, $h = 5$, and $h = 22$ under MSFE, MAE, and QLIKE losses. As in the baseline design, the HAR-type information set is augmented with the same external predictors, and the high-dimensional predictor set used by the machine-learning specifications is screened with the same Elastic Net protocol. All models are estimated separately for each stock and evaluated using the stock's available out-of-sample realized-volatility observations.

The stock-level results are summarized in two steps. First, for each stock, window, horizon, and loss function, we compute aligned full-sample losses and rank the competing models. We then summarize the cross-section using win shares, average and median ranks, MCS inclusion shares, and loss ratios relative to HAR and to the best econometric benchmark. Second, we conduct ticker-level Diebold–Mariano comparisons for the machine-learning models relative to HAR and relative to the best econometric benchmark. The full average-rank, win-count, MCS-inclusion, loss-ratio, DM-summary, and ticker-by-ticker tables are reported in the Online Appendix. Table 10 provides a compact main-text summary focused on the two leading econometric specifications, MSHAR and ARFIMA, together with the cross-sectional DM evidence for the best-performing machine-learning specification.

Several findings emerge. First, the stock-level results strongly reinforce the main S&P 500 evidence: the best-performing specifications are concentrated among the nonlinear and long-memory econometric models rather than among the generic machine-learning architectures. At the one-day horizon, MSHAR is the dominant model across individual equities. In Table 10, MSHAR wins essentially all one-day comparisons across both estimation windows and all three loss functions, while ARFIMA records no one-day wins. MSHAR also has median rank one at $h = 1$ in every window–loss combination and belongs to the MCS for all, or virtually all, stocks at the one-day horizon.

Second, the horizon-specific pattern that appears in the market-index analysis is also clear in the expanded stock-level evidence. MSHAR remains the leading specification at the five-day horizon: its win share exceeds that of ARFIMA under both window schemes and all three loss functions, and its median rank remains one. However, ARFIMA becomes the leading model at the monthly horizon. At $h = 22$, ARFIMA has median rank one in every window–loss combination and accounts for the large majority of stock-level wins. Its long-horizon win share is especially high under MAE and QLIKE, where it wins close to or above 90 percent of the stock-level comparisons in both expanding and rolling windows. This pattern supports the

Table 10: Compact cross-sectional stock-level summary

Window	Loss	$h = 1$	$h = 5$	$h = 22$
Panel A: Win share, MSHAR / ARFIMA				
	MSFE	97.5% / 0.0%	57.5% / 42.5%	20.0% / 80.0%
Expanding	MAE	100.0% / 0.0%	75.0% / 25.0%	12.5% / 87.5%
	QLIKE	100.0% / 0.0%	87.5% / 12.5%	10.0% / 90.0%
	MSFE	100.0% / 0.0%	60.0% / 40.0%	25.0% / 75.0%
Rolling 5-year	MAE	100.0% / 0.0%	90.0% / 10.0%	5.0% / 95.0%
	QLIKE	100.0% / 0.0%	95.0% / 5.0%	7.5% / 92.5%
Panel B: MCS inclusion share, MSHAR / ARFIMA				
	MSFE	100.0% / 7.5%	100.0% / 77.5%	100.0% / 97.5%
Expanding	MAE	100.0% / 0.0%	100.0% / 70.0%	70.0% / 100.0%
	QLIKE	100.0% / 2.5%	100.0% / 55.0%	95.0% / 97.5%
	MSFE	100.0% / 10.0%	100.0% / 80.0%	97.5% / 97.5%
Rolling 5-year	MAE	100.0% / 0.0%	100.0% / 70.0%	52.5% / 97.5%
	QLIKE	100.0% / 0.0%	100.0% / 60.0%	85.0% / 97.5%
Panel C: Median rank, MSHAR / ARFIMA				
	MSFE	1.000 / 9.000	1.000 / 2.000	2.000 / 1.000
Expanding	MAE	1.000 / 10.000	1.000 / 2.000	2.000 / 1.000
	QLIKE	1.000 / 10.000	1.000 / 2.000	2.000 / 1.000
	MSFE	1.000 / 10.000	1.000 / 2.000	2.000 / 1.000
Rolling 5-year	MAE	1.000 / 10.000	1.000 / 2.000	2.000 / 1.000
	QLIKE	1.000 / 10.000	1.000 / 2.000	2.000 / 1.000
Panel D: DM summary, BEST_ML relative to HAR				
	MSFE	20.0 / 0.0 / 37.5	5.0 / 0.0 / 30.0	47.5 / 2.5 / 15.0
Expanding	MAE	12.5 / 0.0 / 42.5	30.0 / 5.0 / 17.5	65.0 / 17.5 / 17.5
	QLIKE	2.5 / 0.0 / 62.5	5.0 / 0.0 / 52.5	30.0 / 2.5 / 25.0
	MSFE	22.5 / 2.5 / 27.5	7.5 / 0.0 / 20.0	45.0 / 7.5 / 15.0
Rolling 5-year	MAE	25.0 / 10.0 / 30.0	57.5 / 10.0 / 2.5	67.5 / 17.5 / 2.5
	QLIKE	5.0 / 0.0 / 55.0	10.0 / 0.0 / 27.5	35.0 / 2.5 / 17.5
Panel E: DM summary, BEST_ML relative to best econometric benchmark				
	MSFE	0.0 / 0.0 / 90.0	0.0 / 0.0 / 92.5	0.0 / 0.0 / 27.5
Expanding	MAE	0.0 / 0.0 / 97.5	0.0 / 0.0 / 97.5	0.0 / 0.0 / 87.5
	QLIKE	0.0 / 0.0 / 97.5	0.0 / 0.0 / 100.0	0.0 / 0.0 / 50.0
	MSFE	0.0 / 0.0 / 95.0	0.0 / 0.0 / 85.0	0.0 / 0.0 / 37.5
Rolling 5-year	MAE	0.0 / 0.0 / 97.5	0.0 / 0.0 / 97.5	0.0 / 0.0 / 100.0
	QLIKE	0.0 / 0.0 / 100.0	0.0 / 0.0 / 100.0	0.0 / 0.0 / 62.5

Notes: Panels A–C report MSHAR / ARFIMA. Panels A and B report percentages; Panel C reports median cross-sectional ranks, where lower ranks indicate lower loss. Panels D and E use the BEST_ML row from the cross-sectional DM summaries. Each DM cell reports lower average loss / significantly lower loss / significantly higher loss for BEST_ML relative to the indicated benchmark, in percent. The best econometric benchmark is selected separately for each stock, window, horizon, and loss function from HAR, THAR, STHAR, MSHAR, and ARFIMA. The cross-sectional stock summary is based on the 40 individual VOLARE stocks and excludes the S&P 500 from the denominators.

paper's main interpretation: latent regime dependence is most useful for short-horizon volatility dynamics, whereas fractional persistence becomes increasingly important as the forecast horizon lengthens.

Third, the MCS evidence qualifies the simple win-count comparisons in a useful way. MSHAR is included in the MCS for all stocks at $h = 1$ and $h = 5$, indicating that even when another model has the lowest average loss for a particular stock, MSHAR is rarely statistically excluded from the superior set at short and medium horizons. ARFIMA, by contrast, is included in the MCS much less often at $h = 1$, becomes more competitive at $h = 5$, and is included for almost all stocks at $h = 22$. Thus, the MCS results do not merely reproduce the win shares; they show that the econometric frontier shifts with the forecast horizon, from a regime-switching HAR structure at short horizons toward long-memory dynamics at longer horizons.

Fourth, the cross-sectional DM summaries show that the machine-learning models do not displace the leading econometric specifications. Relative to HAR alone, the best machine-learning model sometimes has lower average loss, particularly under MAE at longer horizons. This indicates that the ML specifications can occasionally improve on a basic linear HAR benchmark. However, the stricter comparison is against the best econometric benchmark selected separately for each stock, window, horizon, and loss function from HAR, THAR, STHAR, MSHAR, and ARFIMA.[8] Under this comparison, the best ML model has a lower average loss in none of the reported stock-level cells in Table 10, and it never significantly outperforms the best econometric benchmark. In many cases, the best econometric benchmark significantly outperforms the best ML model. This distinction is important: the relevant conclusion is not simply that ML models fail to beat HAR in every setting, but rather that their apparent gains relative to HAR disappear once the benchmark set includes nonlinear regime-dependent and long-memory econometric alternatives.

In Table 11, we report median loss-ratios across 40 stocks with interquartile ranges under rolling 5-year window estimation using MSFE loss. Results under MAE and QLIKE as well as detailed results are are reported in Online Appendix I. The loss-ratio evidence in Table 11 and the Online Appendix as well as detailed out-of-sample losses by window, horizon, and loss metrics in Online Appendix II confirm that these cross-sectional rankings are not driven only by small numerical differences. Relative to HAR, MSHAR delivers the largest median loss reductions at short horizons, whereas ARFIMA becomes increasingly favorable at the monthly horizon. When losses are scaled by the best econometric benchmark selected separately for each stock, window, horizon, and loss function, the machine-learning specifications do not display systematic improvements. Thus, the magnitude evidence reinforces the same horizon-specific

[8] The best econometric benchmark is not fixed ex ante, but the selected benchmark is almost always MSHAR at $h = 1$ and almost always ARFIMA at $h = 22$.

Table 11: Cross-sectional loss ratios under rolling window using MSFE loss

Model	$h = 1$	$h = 5$	$h = 22$
Panel A: Median loss ratio relative to HAR			
HAR	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]
THAR	1.041 [1.011,1.073]	1.046 [1.021,1.072]	1.036 [1.013,1.072]
STHAR	1.053 [1.018,1.093]	1.039 [1.024,1.082]	1.028 [1.010,1.086]
MSHAR	0.588 [0.498,0.711]	0.634 [0.545,0.727]	0.644 [0.576,0.690]
ARFIMA	1.618 [1.231,2.380]	0.663 [0.609,0.746]	0.478 [0.442,0.551]
XGB	1.347 [1.164,1.579]	1.229 [1.104,1.365]	1.031 [0.994,1.082]
DNN	1.081 [1.034,1.197]	1.113 [1.055,1.193]	1.120 [1.080,1.285]
LSTM	1.347 [1.201,1.580]	1.123 [1.096,1.221]	1.043 [1.011,1.082]
LSTM-A	1.374 [1.202,1.552]	1.128 [1.081,1.219]	1.055 [1.030,1.105]
GRU-A	1.330 [1.224,1.540]	1.110 [1.073,1.191]	1.049 [1.014,1.096]
Panel B: Median loss ratio relative to best econometric benchmark			
HAR	1.701 [1.406,2.007]	1.670 [1.597,1.836]	2.158 [1.903,2.262]
THAR	1.831 [1.473,2.168]	1.774 [1.686,1.969]	2.188 [1.990,2.447]
STHAR	1.869 [1.544,2.151]	1.785 [1.655,1.959]	2.183 [2.005,2.443]
MSHAR	1.000 [1.000,1.000]	1.000 [1.000,1.086]	1.319 [1.008,1.583]
ARFIMA	2.747 [2.175,3.771]	1.051 [1.000,1.364]	1.000 [1.000,1.003]
XGB	2.260 [1.923,2.696]	2.121 [1.937,2.431]	2.171 [1.965,2.418]
DNN	1.936 [1.518,2.490]	1.891 [1.717,2.174]	2.386 [2.125,2.808]
LSTM	2.299 [1.939,2.720]	2.007 [1.822,2.149]	2.216 [2.051,2.403]
LSTM-A	2.299 [1.988,2.736]	1.980 [1.871,2.171]	2.247 [2.032,2.412]
GRU-A	2.272 [2.007,2.721]	1.924 [1.794,2.126]	2.191 [2.025,2.403]

Notes: Entries report median loss ratios across stocks; interquartile ranges are in brackets. Ratios below one indicate lower loss than the corresponding benchmark. The best econometric benchmark is selected separately for each ticker, horizon, window, and metric from HAR, THAR, STHAR, MSHAR, and ARFIMA.

interpretation as the win-count, MCS, and DM summaries: latent regime dependence is most useful at short horizons, while long-memory persistence is more important at longer horizons.

The matched-sample S&P 500 results in the Online Appendix provide an additional check on the interpretation. Because those results use the same 2015–2025 sample span and the same 5-year rolling-window design as the individual stocks, they help separate two effects that are otherwise confounded: the effect of moving from a long market-index sample to a shorter common sample, and the effect of moving from the aggregate index to individual equities. The appendix evidence shows whether the short-horizon MSHAR and long-horizon ARFIMA pattern survives when the S&P 500 is placed on the same sample footing as the stock-level exercise. We therefore use the matched-sample S&P 500 rows as a bridge between the main long-sample S&P 500 analysis and the 40-stock cross-sectional analysis, while keeping the formal stock cross-sectional summaries based only on individual equities.

Overall, the 40-stock exercise provides a stronger and more transparent robustness check than the earlier selected-stock analysis. It removes the ad hoc nature of selecting a small number of firms and evaluates the model rankings across the full individual-equity universe available in the VOLARE stock file. At the same time, we interpret the results cautiously: the VOLARE stock universe is a data-availability-based set of liquid U.S. equities rather than a random sample of all listed firms. Within that universe, however, the evidence is clear. Moving from

the S&P 500 to individual equities does not make the more flexible machine-learning models systematically dominant. Instead, the stock-level evidence reinforces the central conclusion of the paper: the most useful forms of complexity are economically structured forms of volatility dependence—latent regimes at short horizons and long memory at longer horizons—rather than generic nonlinear prediction architectures.

6 Conclusion

This paper examines which representations of persistence and nonlinearity generate reliable out-of-sample gains in realized-volatility forecasting and whether machine learning adds value beyond econometric models designed to capture long memory and regime dependence. We compare HAR, ARFIMA, THAR, STHAR, and MSHAR with XGBoost and several neural-network architectures. The analysis covers a long S&P 500 sample and 40 individual U.S. equities from VOLARE, using both realized-volatility history and an extended predictor set selected by Elastic Net. Performance is evaluated using statistical losses, Diebold–Mariano tests, Model Confidence Sets, realized utility, and VaR and expected-shortfall criteria.

The central finding is that forecast performance depends less on generic nonlinear flexibility than on representing the volatility feature most relevant at a given horizon. MSHAR performs most reliably at the one-day horizon, indicating that latent regime dependence contains useful short-run predictive information. ARFIMA generally leads at the twenty-two-day horizon, highlighting the importance of fractional persistence, while results at the five-day horizon are more mixed. Thus, targeted representations of regime dependence and long memory are generally more effective than generic flexibility, with their relative value varying systematically across horizons.

These rankings are broadly stable across expanding and rolling windows and across MSFE, MAE, and QLIKE. HAR remains a useful and robust benchmark, but it does not define the econometric forecasting frontier once ARFIMA and nonlinear HAR specifications are considered. ML models sometimes outperform HAR, but rarely displace the leading econometric models. This distinction helps reconcile the apparent success of ML relative to HAR with its weaker performance against a broader set of econometric benchmarks.

Expanding the information set does not materially alter this conclusion. Macro-financial, uncertainty, sentiment, geopolitical-risk, factor, and return-asymmetry variables improve some model–horizon combinations, but the gains are not sufficiently broad or stable to change the overall rankings. The economic-value and tail-risk exercises yield less uniform results but broadly reinforce the statistical evidence: realized utility preserves the horizon-dependent pattern, and VaR and expected-shortfall comparisons provide no systematic evidence that ML outperforms the leading econometric specifications.

The cross-sectional results show that the main findings are not peculiar to aggregate S&P 500 volatility. Across 40 individual equities, MSHAR performs most strongly at the short horizon, whereas ARFIMA becomes increasingly competitive and generally leads at the monthly horizon. The best ML specification does not systematically outperform the best econometric model across stocks, horizons, windows, and loss functions. Results for the matched-sample S&P 500 further indicate that this pattern is not explained solely by shorter equity histories or 5-year rolling windows.

A range of robustness exercises supports these conclusions. The principal rankings remain intact under quarterly re-estimation, a log realized-volatility target, and a realized-kernel measure. Neural-network diagnostics and additional random initializations suggest that overfitting or initialization instability is unlikely to explain the ML rankings. Moreover, comparisons with corresponding TAQ measures show that the VOLARE series have similar distributional and persistence properties and are highly correlated with TAQ, reducing concerns that the equity results reflect peculiarities of the data source.

These findings should not be interpreted as evidence that MSHAR or ARFIMA represents the true volatility-generating process, or that ML is inherently unsuitable for volatility forecasting. Performance is conditional on the samples, predictors, architectures, and asset-by-asset design considered here. ML may deliver larger gains with broader panels or richer order-book, textual, options-based, or other high-dimensional information. Nevertheless, such gains should be evaluated against econometric benchmarks that already capture the principal time-series features of realized volatility, rather than against HAR alone.

The main implication is that interpretability and forecast accuracy need not conflict. Carefully structured econometric models can remain competitive with, and often outperform, more flexible and computationally intensive alternatives. Useful structure is also horizon dependent: latent regime variation is especially informative in the short run, whereas fractional persistence matters more at longer horizons. Future research could use controlled simulations to identify when ML methods recover long-memory and regime-switching dynamics, develop hybrid models combining these structures with flexible architectures, and examine whether pooling information across assets allows ML to exploit common volatility dynamics more effectively.

References

Alexander, Carol (2009). Market Risk Analysis, Value at Risk Models. John Wiley and Sons.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens (2001). The Distribution of Realized Stock Return Volatility, *Journal of Financial Economics* 61: 43–76.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys (2003). Modeling and Forecasting Realized Volatility. *Econometrica* 71: 579–625.

Aruoba, S.B., Diebold, F.X. and Scotti, C. (2009). Real-Time Measurement of Business Conditions, *Journal of Business and Economic Statistics* 27:4, pp. 417-27.

Audrino, F., and Knaus, S. D. (2016). Lassoing the HAR Model: A Model Selection Perspective on Realized Volatility Dynamics, *Econometric Reviews,* 35, 1485–1521.

Audrino, F., Sigrist, F., and Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility, *International Journal of Forecasting,* 36, 334–357.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR).*

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics, *Journal of Econometrics* 73: 5–59.

Baillie, R. T., Calonaci, F., Cho, D., and Rho, S. (2019). Long memory, realized volatility and heterogeneous autoregressive models. *Journal of Time Series Analysis*, 40, 609–628.

Baker, S. R., N. Bloom, S. Davis, J. Steven (2016). Measuring economic policy uncertainty. *Q. J. Econ.* 131 (4), 1593–1636

Barkan, O, J. Benchimo, I. Caspi, E. Cohen, A. Hammer, and N. Koenigstein (2023). Forecasting CPI inflation components with Hierarchical Recurrent Neural Networks, *International Journal of Forecasting* 39. 1145-1162.

Barndorff-Nielsen, Ole E., and Neil Shephard (2002). Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64: 253-280.

Barndorff-Nielsen, Ole E., Peter Reinhard Hansen, Asger Lunde, and Neil Shephard (2008). Designing Realized Kernels to Measure the Ex Post Variation of Equity Prices in the Presence of Noise, *Econometrica* 76(6): 1481–1536.

Barndorff-Nielsen, Ole E., Peter Reinhard Hansen, Asger Lunde, and Neil Shephard (2009). Realized Kernels in Practice: Trades and Quotes, *The Econometrics Journal* 12(3): C1–C32.

Barone-Adesi, G., K. Giannopoulos, and L. Vosper (1999). Var without correlations for portfolios of derivative securities. *J. Futures Markets* 19 (5), 583–602.

Bauwens, L., Hafner, C., and Laurent, S. (2012). *Volatility Models and Their Applications*, Volume 15. John Wiley and Sons, Inc.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroscedasticity. Journal of Econometrics 31: 307-327.

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

Brownlees, Christian, and Giampiero Gallo (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns, Computational Statistics and Data Analysis 51, 2232-2245.

Branco, R.R., A. Rubesam, M. Zevallos (2024). Forecasting realized volatility: Does anything beat linear models?, *Journal of Empirical Finance*, 78, 1-22.

Bucci, A. (2020) Realized Volatility Forecasting with Neural Networks *Journal of Financial Econometrics* 18: 502–531.

Buckman, S.R, A. H. Shapiro, M. Sudhof, and D. Wilson (2020). News sentiment in the time of COVID-19. *FRBSF Econ. Lett.* 8 (1), 5–10.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* 16: 1190–1208.

Carrasco, M., Hu, L., and Ploberger, W. (2014). Optimal test for Markov switching parameters. *Econometrica*, 82(2), 765–784.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 785-794).

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734.

Christensen, K., M. Siggaard, and B. Veliyev (2023). A Machine Learning Approach to Volatility Forecasting, *Journal of Financial Econometrics,* Vol. 21. 1680-1727.

Christoffersen, P. (1998). Evaluating Interval Forecasts *International Economic Review* Vol. 39, pp. 841–862.

Christoffersen, P. and D. Pelletier (2004). Backtesting value-at-risk: A duration-based approach. *J. Empir. Finance* 2, 84–108.

Chung, J., C. Gulcehre, K. Cho, and Y. Bengio (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv, https://arxiv.org/abs/1412.3555.

Cipollini, F., Cruciani, G., Gallo, G. M., Insana, A., Otranto, E., and Spagnolo, F. (2026). VOLatility Archive for Realized Estimates (VOLARE). arXiv:2602.19732 [q-fin.ST]. https://doi.org/10.48550/arXiv.2602.19732.

Corsi, F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility, *Journal of Financial Econometrics* 7: 174–196.

Corsi, Fulvio, Renó, Roberto (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *J. Busines and Economic. Statistics* 30 (3), 368–380.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function *Math. Control Signal Systems* 2, 303–314.

Díaz J.D., E. Hansen, G. Cabrera (2024). Machine-learning stock market volatility: Predictability, drivers, and economic value, *International Review of Financial Analysis,* 94 103286.

Dey, R., and Salemt, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international midwest symposium on circuits and systems (pp. 1597–1600). IEEE.

Diebold, F.X. and R.S. Mariano. (1995). Comparing Predictive Accuracy, *Journal of Business and Economic Statistics,* 13: 253-63

Donaldson, G. R., and M. Kamstra (1997). An Artificial Neural network-GARCH Model for International Stock Return Volatility, *Journal of Empirical Finance* 4: 17–46.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica 50: 987-1007.

Fama, E., and K. French. (1993). Common Risk Factors in the Returns on Stocks and Bonds, *Journal of Financial Economics* 33: 3–56.

Fernandes, M., M. C. Medeiros, and M. Scharth (2014). Modeling and Predicting the CBOE Market Volatility Index, *Journal of Banking and Finance* 40: 1–10.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4), 1680–1707. *Journal of Econometrics*, 211(2), 388–413.

Franses, P. H. and D. van Dijk (2000). Non-Linear Time Series Models in Empirical Finance. Cambridge University Press.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics* Vol. 29, pp. 1189-1232.

Giot, Pierre, Laurent, S., (2004). Modelling daily value-at-risk using realized volatility and ARCH type models. *J. Empir. Finance* 11 (3), 379–398.

Goldfeld, S. M., and R. E. Quandt (1973). A Markov model for switching regressions, *Journal of Econometrics* 1: 3–15. https://doi.org/10.1016/0304-4076(73)90002-X.

Gonzalez-Rivera, G., T.-H, Lee, and S. Mishra (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood, *International Journal of Forecasting,* 20(4), 629–645.

Gonzalo, J., and J.-Y. Pitarakis (2002). Estimation and model selection based inference in single and multiple threshold models. Journal of Econometrics 110: 319–352. https://doi.org/10.1016/S0304-4076(02)00098-2

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning, MIT Press, http://www.deeplearningbook.org.

Granger, C. W. J., and R. Joyeux (1980). An Introduction to Long-Memory Time Series Models and Fractional Differencing. *Journal of Time Series Analysis* 1: 15–29.

Granger, Clive W. J. and T. Terasvirta (1993). Modelling Non-Linear Economic Relationships, Oxford University Press, New York.

Graves, A., Mohamed, A., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In ICASSP 2013 , pages 6645–6649

Graves, A. (2013). Generating sequences with recurrent neural networks. Technical report, arXiv:1308.0850.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In ICML 2014

Gunnarsson, E. S., Isern, H. R., Kaloudis, A., Risstad, M., Vigdel, B., and Westgaard, S. (2024). Prediction of realized volatility and implied volatility indices using AI and machine learning: A review, *International Review of Financial Analysis,* 93, 103221.

Hamid, S. A., and Z. Iqbal (2004). Using Neural Networks for Forecasting Volatility of SandP 500 Index Futures Prices, *Journal of Business Research* 57: 1116–1125.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* 57: 357–384. https://doi.org/10.2307/1912559.

Hamilton, J. D. (1994). Time Series Analysis. Princeton, NJ: Princeton University Press.

Hansen, B. E. (1996). Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis. *Econometrica* 64: 413–430.

Hansen, B. E. (2000). Sample splitting and threshold estimation. Econometrica 68: 575–603. https://doi.org/10.1111/1468-0262.00124.

Hansen, B. E. (2011). Threshold autoregression in economics. Statistics And Its Interface 4: 123–127. https://doi.org/10.4310/SII.2011.v4.n2.a4

Hansen, P. R., and A. Lunde (2005). A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics* 20: 873–889.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science and Business Media.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778).

Hill, Tim, Marcus O'Connor, and William Remus (1996). Neural Network Models for Time Series Forecasts *Management Science* 42: 1082–1092.

Hillebrand, E. and M. C. Medeiros (2010). The Benefits of Bagging for Forecast Models of Realized Volatility. *Econometric Reviews* 29 (5-6), 571-593.

Holló, Dániel, Manfred Kremer, and Marco Lo Duca (2012). CISS – A Composite Indicator of Systemic Stress in the Financial System. *ECB Working Paper* No. 1426, March.

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735-1780.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). Multilayer Feedforward Networks Are Universal Approximators *Neural Networks* 2: 359–366.

Hosking, J. R. M. (1981). Fractional Differencing. *Biometrika* 68: 165–176.

Hu, Y. M., and C. Tsoukalas (1999). Combining Conditional Volatility Forecasts Using Neural Networks: An Application to the EMS Exchange Rates, *Journal of International Financial Markets, Institutions and Money* 9: 407–422.

Iacoviello, Matteo and Jonathan Tong (2026). The AI-GPR Index: Measuring Geopolitical Risk using Artificial Intelligence. Working paper, April.

Izzeldin, Marwan, M. Kabir Hassan, Vasileios Pappas, and Mike Tsionas (2019). Forecasting Realised Volatility Using ARFIMA and HAR Models, *Quantitative Finance* 19: 1627–1638.

Karpathy, A. (2015). The unreasonable effectiveness of recurrent neural networksm https://karpathy.github.io/2015/05/21/rnn-effectiveness/

Kılıç (2011). Long Memory and Nonlinearity in Conditional Variances: A Smooth Transition FIGARCH Model, *Journal of Empirical Finance* 18, 368–378.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching, *Journal of Econometrics* 60: 1–22.

Kingma, Diederik P., and Jimmy Ba (2014). Adam: A Method for Stochastic Optimization, Working paper.

Koenker, R., and G. Bassett (1978). Regression Quantiles, *Econometrica* 46: 33–50.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1097-1105).

Kupiec, P. (1995). Techniques for Verifying the Accuracy of Risk Management Models *Journal of Derivatives* Vol. 3, pp. 73 – 84.

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, Vol. 30, pp. 6402–6413.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**(11), 2278-2324.

Leybourne, S., P. Newbold and D. Vougas (1998) Unit Roots and Smooth Transitions *Journal of Time Series Analysis,* 19, 83-97.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard (2015). Does Anything Beat 5-Minute RV? A Comparison of Realized Measures across Multiple Asset Classes, *Journal of Econometrics* 187: 293–311.

Liu, F., Pantelous, A. A., and von Mettenheim, H. J. (2018). Forecasting and trading high frequency volatility on large indices, *Quantitative Finance,* 18, 737–748.

Lipton, Z. C. and J. Berkowitz, and C. Elkan (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning, arXivhttps://arxiv.org/abs/1506.00019

Lundberg, Scott M. and Su-In Lee (2017). A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems* 30.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1412-1421).

Maas, A. L., A. Y. Hannun, and A. Y. Ng (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the 30th International Conference on Machine Learning (ICML) Workshops*, Vol. 30.

Masters, T. (1993). Practical neural network recipes in $C++$, New York: Academic Press.

Mittnik, S., Robinzonov, N., and Spindler, M. (2015). Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of Banking and Finance*, 58, 1–14.

Motegi, K. C. Xiaojing, Hamori, S. and X. Haifeng (2020). Moving average threshold heterogeneous autoregressive (MAT-HAR) models, *Journal of Forecasting* 39:1035-1042.

Newey, W. K., and K. D. West (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55: 703–708.

Patton, Andrew J., and Kevin Sheppard (2009). Evaluating Volatility and Correlation Forecasts. In T. Mikosch, J.-P. KreiB, R. A. Davis, and T. G. Andersen (eds.), *Handbook of Financial Time Series*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 801–838.

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

Patton, Andrew J. and Kevin Sheppard (2015). Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility, *Review of Economics and Statistics* 97: 683–697.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk).

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, Vol. 32, pp. 8024–8035.

Prechelt, L. (1998). Early Stopping — But When? In G. B. Orr and K.-R. Müller (eds.), *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, Vol. 1524. Springer, Berlin, pp. 55–69.

Quandt, R. E. (1972). A new approach to estimating switching regressions, *Journal of the American Statistical Association* 67: 306–310. https://doi.org/10.1080/01621459.1972.10482378.

Rahimika, E. and S-H Poon (2024). Machine Learning for Realised Volatility Forecasting, Working Paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3707796&download=yes.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, **323**(6088), 533-536.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research* 15 (1), 1929-1958.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In NIPS 2014, arXiv:1409.3215.

Takahashi, M., Omori, Y., and Watanabe, T. (2023). *Stochastic Volatility and Realized Stochastic Volatility Models,* SpringerBriefs in Statistics. Springer Nature.

Taylor, S. J. (1982). Financial Returns Modeled by the Product of Two Stochastic Processes - a Study of the Daily Sugar Prices 1961-75, In O. D. Anderson (ed.), Time Series Analysis: Theory and Practice. Amsterdam: North-Holland, 203-226.

Terasvirta, T. (1994). Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models, *Journal of the American Statistical Association,* 89, 208-218.

van Dijk, D., T. Teräsvirta, and P. H. Franses (2002). Smooth Transition Autoregressive Models — A Survey of Recent Developments. *Econometric Reviews* 21: 1–47.

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., and Hinton, G. (2014). Grammar as a foreign language. Technical report, arXiv:1412.7449.

Vortelinos, D. I. (2017). Forecasting Realized Volatility: HAR against Principal Components Combining, Neural Networks and GARCH, *Research in International Business and Finance* 39, 824–839.

Wang J, Ma F, Liang C, Chen Z. (2022). Volatility forecasting revisited using Markov-switching with time-varying probability transition, *Int J Fin Econ*.27:1387–1400. https://doi.org/10.1002/ijfe.2221

Wong, Z. Y, Chin, W. C, Tan, S. H. (2016). Daily value-at-risk modeling and forecast evaluation: The realized volatility approach. *J. Finance Data Sci.* 2 (3), 171–187.

XGBoost Developers (2025). XGBoost Python API Reference, *XGBoost Documentation*. Available at: `https://xgboost.readthedocs.io/`.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visualattention. In ICML'2015, arXiv:1502.03044.

Zhang, Peter G. (2003). Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model *Neurocomputing* 50: 159–175.

Zhang, C., Zhang, Y., Cucuringu, M., and Qian, Z. (2023). Volatility Forecasting with Machine Learning and Intraday Commonality. *Journal of Financial Econometrics,* 1–39.

Zhu, H., L. Bai, L. He, and Z. Liu (2023). Forecasting realized volatility with machine learning: Panel data perspective. *Journal of Empirical Finance* 73, 251-271.

Zou, H., and T. Hastie (2005). Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67: 301–320.

Appendices

A Robustness checks

A.1 Robustness to more frequent re-estimation under extended information set

A natural concern is that the relatively weaker performance of some machine-learning models may reflect infrequent parameter updating rather than limitations of the models themselves. To address this issue, we re-estimate all models at a quarterly frequency while keeping the rolling-window design, forecast horizons, predictor set, and Elastic Net screening protocol unchanged. This exercise is intended to isolate the role of more frequent model updating relative to the baseline yearly retraining scheme.

Table A1: Rolling-window robustness check under extended EN-screened predictor set: average QLIKE by retraining frequency

Horizon	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel A: Yearly retraining										
$h = 1$	0.583	0.584	0.585	**0.575**	0.760	0.585	0.586	0.585	0.586	0.584
$h = 5$	0.623	0.627	0.627	0.601	**0.596**	0.633	0.629	0.630	0.631	0.629
$h = 22$	0.674	0.678	0.678	0.624	**0.593**	0.685	0.699	0.698	0.711	0.699
Panel B: Quarterly retraining										
$h = 1$	0.580	0.581	0.581	**0.575**	1.667	0.583	0.583	0.584	0.583	0.582
$h = 5$	0.624	0.629	0.628	0.601	**0.594**	0.633	0.628	0.631	0.628	0.627
$h = 22$	0.681	0.687	0.687	0.632	**0.594**	0.682	0.703	0.709	0.725	0.721

Notes: Entries report the average rows from the rolling-window QLIKE tables under extended-information set of predictors with Elastic Net screening. Panel A uses yearly retraining; Panel B uses quarterly retraining. Lower values indicate better forecasting performance. Bold-underlined entries denote the lowest average QLIKE within each row. The full yearly and quarterly tables for MSFE, MAE, and QLIKE are reported in the Online Appendix.

Table A1 summarizes the average rolling-window QLIKE results under yearly and quarterly retraining.[9] The most important finding is that increasing the retraining frequency does not materially alter the qualitative ordering of the leading models. At the one-day horizon, MSHAR remains the best-performing specification under both yearly and quarterly updating, with average QLIKE equal to 0.575 in both cases. At the five-day horizon, ARFIMA remains the best model, and its average QLIKE improves slightly from 0.596 under yearly retraining to 0.594 under quarterly retraining. At the twenty-two-day horizon, ARFIMA again remains the best specification, with average QLIKE essentially unchanged at 0.593–0.594. Thus, the dominant conclusion from the baseline specification—namely that the strongest performance

[9] Additional results are provided in online Appendix F.

remains concentrated in the nonlinear econometric models at short horizons and in ARFIMA at longer horizons—is robust to more frequent re-estimation.

The machine-learning models do not exhibit the kind of broad-based improvement that would be needed to overturn the baseline ranking. At the short horizon, quarterly retraining produces only very small changes in average QLIKE for MLP, LSTM, LSTM-A, and GRU-A, and none of these changes is large enough to challenge the leading performance of MSHAR. At the medium horizon, a few machine-learning specifications improve marginally, but the overall ranking remains essentially unchanged and the best average performance continues to belong to ARFIMA. At the long horizon, more frequent retraining likewise does not generate a systematic improvement in relative machine-learning performance; if anything, the longer-horizon recurrent-network results remain slightly weaker than the leading econometric alternatives.

The broader quarterly results reported in the Online Appendix point to the same conclusion. Across MSFE, MAE, and QLIKE, more frequent quarterly retraining produces only modest changes in loss magnitudes and limited reshuffling among the middle-ranked models, while leaving the identity of the strongest models largely intact. In particular, the quarterly results do not suggest that the comparatively weaker performance of the machine-learning models in the baseline exercise is mainly driven by yearly retraining. Rather, the evidence indicates that the main performance differences across models are robust to increasing the frequency of re-estimation.

Taken together, these findings imply that retraining frequency is not a first-order driver of the main empirical conclusions. Quarterly re-estimation provides a useful robustness check, but it does not materially change the relative standing of the main econometric and machine-learning models under the extended information set.

A.2 Robustness to using log realized volatility

In order to assess whether the main conclusions depend on the scale of the forecasting target, we re-estimate models using log realized volatility as the target. We re-ran the rolling extended-information set exercise using log realized volatility while keeping the information set, forecast horizons, predictor-screening protocol, and yearly re-estimation design unchanged. Table A2 compares the average rolling-window results under the baseline level target and the log-transformed target.

The broad conclusions are robust to this transformation. At the one-day horizon, MSHAR remains the best-performing specification under all three loss functions in both the level and log-target exercises. At the twenty-two-day horizon, ARFIMA continues to dominate under MSFE and QLIKE, while under MAE the comparison between ARFIMA and MSHAR becomes very tight, with MSHAR only marginally ahead under the log target. Thus, the main qualitative pattern of the paper—namely that nonlinear econometric models dominate at short

horizons and ARFIMA remains especially strong at longer horizons—is preserved under the log specification.

Table A2: Rolling-window robustness check under extended EN-screened set of predictors: average losses for level and log realized volatility targets

Horizon	Target	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel A: MSFE											
$h = 1$	Level	0.086	0.085	0.086	**0.053**	0.156	0.104	0.082	0.100	0.100	0.093
	Log	0.085	0.085	0.086	**0.063**	0.281	0.118	0.089	0.088	0.086	0.088
$h = 5$	Level	0.166	0.173	0.174	**0.099**	0.103	0.210	0.176	0.193	0.194	0.188
	Log	0.167	0.167	0.168	0.104	**0.102**	0.216	0.182	0.188	0.197	0.191
$h = 22$	Level	0.254	0.265	0.264	0.155	**0.104**	0.295	0.268	0.272	0.277	0.272
	Log	0.253	0.264	0.263	0.166	**0.135**	0.286	0.280	0.290	0.291	0.267
Panel B: MAE											
$h = 1$	Level	0.172	0.174	0.174	**0.151**	0.262	0.180	0.181	0.182	0.183	0.177
	Log	0.168	0.169	0.170	**0.155**	0.224	0.179	0.170	0.170	0.170	0.169
$h = 5$	Level	0.240	0.245	0.244	0.201	**0.193**	0.259	0.255	0.249	0.250	0.248
	Log	0.227	0.229	0.229	**0.161**	0.184	0.247	0.234	0.236	0.239	0.239
$h = 22$	Level	0.303	0.314	0.313	0.241	**0.191**	0.321	0.311	0.299	0.303	0.303
	Log	0.277	0.289	0.289	**0.194**	0.196	0.301	0.295	0.289	0.295	0.286
Panel C: QLIKE											
$h = 1$	Level	0.583	0.584	0.585	**0.575**	0.760	0.585	0.586	0.585	0.586	0.584
	Log	0.582	0.581	0.582	**0.573**	0.599	0.586	0.582	0.583	0.583	0.582
$h = 5$	Level	0.623	0.627	0.627	0.601	**0.596**	0.633	0.629	0.630	0.631	0.629
	Log	0.623	0.625	0.624	**0.582**	0.590	0.635	0.627	0.629	0.630	0.628
$h = 22$	Level	0.674	0.678	0.678	0.624	**0.593**	0.685	0.699	0.698	0.711	0.699
	Log	0.677	0.682	0.683	0.603	**0.595**	0.691	0.684	0.693	0.691	0.677

Notes: Entries report the average rows from the rolling-window extended-information set performance tables with yearly retraining/re-estimation. "Level" denotes the baseline specification using realized volatility in levels, while "Log" denotes the robustness check using log realized volatility as the forecasting target. Lower values indicate better performance. Bold entries indicate the lowest average loss within each horizon-target row. The full yearly and episode tables under the log-target specification are reported in the Online Appendix.

The main difference appears at the intermediate horizon. Under the baseline level target, MSHAR is best under MSFE at $h = 5$, whereas ARFIMA is best under MAE and QLIKE. Under the log target, this pattern shifts: ARFIMA becomes the best model under MSFE, while MSHAR becomes the best model under MAE and QLIKE. This change is economically modest but indicates that the medium-horizon comparison between ARFIMA and MSHAR is somewhat sensitive to the loss function once the target is log transformed. Importantly, however, the shift occurs within the set of leading econometric models and does not alter the broader ranking of model classes.

The machine-learning models do not exhibit a systematic improvement large enough to overturn the baseline ordering. Their average losses move only modestly under the log transformation, and none of the machine-learning specifications becomes the best overall performer at any horizon. Accordingly, the log-target robustness check does not support the view that the baseline findings are an artifact of forecasting realized volatility in levels rather than logs. In-

stead, it suggests that the principal conclusions are stable, with only limited reshuffling between MSHAR and ARFIMA at the five-day horizon.[10]

A.3 Robustness to alternative realized-volatility measure

A natural question is whether the model rankings depend on the particular realized-volatility proxy used in the analysis. A consistently constructed realized-kernel series is not available for the S&P 500 over the full 1996–2025 sample used in the main index-level exercise. We therefore conduct the alternative-measure robustness analysis for the 40 individual equities, for which VOLARE provides both the baseline five-minute realized-variance measure and a realized-kernel measure over a common sample. Although this exercise does not provide an alternative-measure comparison for the long-sample S&P 500 series, it offers a broad cross-sectional assessment of whether the relative performance of the competing model classes is sensitive to realized-volatility measurement. Specifically, we repeat the stock-level cross-sectional exercise using the realized kernel as the forecasting target.[11]

Table A3 summarizes the realized-kernel results. The qualitative message is unchanged.[12] At the one-day horizon, MSHAR remains the dominant specification across all three loss functions. Its win share is 97.5% under MSFE and 100.0% under MAE and QLIKE, and it is included in the 5% Model Confidence Set for all stocks and all loss functions. The median rank of MSHAR is one under each loss, while ARFIMA performs poorly at this short horizon, with median rank equal to ten under all three losses. This evidence is consistent with the baseline conclusion that short-horizon stock volatility forecasts benefit most from allowing latent regime dependence in the HAR dynamics.

The intermediate horizon continues to favor MSHAR, although ARFIMA becomes more competitive. At $h = 5$, MSHAR has win shares of 95.0%, 92.5%, and 97.5% under MSFE, MAE, and QLIKE, respectively, and is included in the MCS for all stocks under each loss. ARFIMA obtains lower loss for a nontrivial subset of stocks at this horizon and has MCS inclusion

[10] The full rolling-window yearly and episode tables under the log-target specification are reported in online Appendix G.

[11] The realized kernel is a high-frequency estimator of the ex post quadratic variation of prices that is designed to remain robust in the presence of market microstructure noise. Unlike simple realized variance, which sums squared intraday returns and can become biased when sampling too finely, the realized-kernel estimator combines intraday return autocovariances using kernel weights,

$$RK_t = \gamma_{0,t} + \sum_{j=1}^{H} k\left(\frac{j}{H+1}\right)(\gamma_{j,t} + \gamma_{-j,t}),$$

where $\gamma_{j,t}$ denotes the j-th realized autocovariance of high-frequency returns on day t, $k(\cdot)$ is a kernel function, and H is the bandwidth. The weighting of autocovariances provides a noise-robust estimate of daily integrated variance (Barndorff-Nielsen et al., 2008, 2009).

[12] For brevity, we do not report full detailed tables under realized kernel-based measure. These detailed results with ticker-by-ticker out-sample loss measures for all models with MCS test results are available upon request.

Table A3: Compact cross-sectional stock-level summary: realized-kernel robustness

Window	Loss	$h = 1$	$h = 5$	$h = 22$
Panel A: Win share, MSHAR / ARFIMA				
	MSFE	97.5% / 0.0%	95.0% / 40.0%	85.0% / 100.0%
Rolling 5-year	MAE	100.0% / 0.0%	92.5% / 60.0%	85.0% / 100.0%
	QLIKE	100.0% / 0.0%	97.5% / 20.0%	85.0% / 100.0%
Panel B: MCS inclusion share, MSHAR / ARFIMA				
	MSFE	100.0% / 20.0%	100.0% / 80.0%	100.0% / 100.0%
Rolling 5-year	MAE	100.0% / 0.0%	100.0% / 80.0%	90.0% / 100.0%
	QLIKE	100.0% / 0.0%	100.0% / 60.0%	97.5% / 100.0%
Panel C: Median rank, MSHAR / ARFIMA				
	MSFE	1.000 / 10.000	1.000 / 2.000	1.000 / 1.000
Rolling 5-year	MAE	1.000 / 10.000	1.000 / 1.000	1.000 / 1.000
	QLIKE	1.000 / 10.000	1.000 / 2.000	1.000 / 1.000
Panel D: DM summary, BEST_ML relative to HAR				
	MSFE	12.5 / 2.5 / 52.5	7.5 / 0.0 / 20.0	42.5 / 5.0 / 7.5
Rolling 5-year	MAE	20.0 / 2.5 / 57.5	47.5 / 5.0 / 2.5	65.0 / 27.5 / 2.5
	QLIKE	7.5 / 0.0 / 70.0	15.0 / 0.0 / 35.0	32.5 / 5.0 / 12.5
Panel E: DM summary, BEST_ML relative to best econometric benchmark				
	MSFE	0.0 / 0.0 / 95.0	0.0 / 0.0 / 95.0	2.5 / 0.0 / 52.5
Rolling 5-year	MAE	0.0 / 0.0 / 97.5	0.0 / 0.0 / 97.5	2.5 / 2.5 / 80.0
	QLIKE	0.0 / 0.0 / 100.0	0.0 / 0.0 / 100.0	2.5 / 0.0 / 57.5

Notes: The table reports the stock-level robustness exercise in which realized volatility is measured using the realized-kernel variable rather than the baseline 5-minute realized-variance measure. Panels A–C report MSHAR / ARFIMA. Panels A and B report percentages; Panel C reports median cross-sectional ranks, where lower ranks indicate lower loss. Panels D and E use the BEST_ML row from the cross-sectional DM summaries. Each DM cell reports lower average loss / significantly lower loss / significantly higher loss for BEST_ML relative to the indicated benchmark, in percent. The best econometric benchmark is selected separately for each stock, window, horizon, and loss function from HAR, THAR, STHAR, MSHAR, and ARFIMA. The cross-sectional stock summary is based on the 40 individual VOLARE stocks and excludes the S&P 500 from the denominators.

shares of 80.0%, 80.0%, and 60.0% under MSFE, MAE, and QLIKE. Thus, the realized-kernel exercise reproduces the same transition documented in the baseline results: regime-switching HAR dynamics dominate at short horizons, while fractional long-memory dynamics become increasingly relevant as the forecast horizon lengthens.

At the monthly horizon, the evidence points to a close role for MSHAR and ARFIMA. ARFIMA has median rank one and 100.0% MCS inclusion under all three losses, while MSHAR also has median rank one and remains in the MCS for 90.0% to 100.0% of stocks depending on the loss function. The detailed loss-ratio tables reinforce this interpretation. Relative to HAR, MSHAR has median loss ratios below one at all horizons and under all three losses. ARFIMA is worse than HAR at $h = 1$ but improves on HAR at $h = 5$ and $h = 22$, with particularly large gains at the monthly horizon. For example, as can be seen in Table A4, under MSFE the median loss ratio relative to HAR is 0.620 for MSHAR at $h = 1$, while ARFIMA's corresponding ratio is 1.558; at $h = 22$, the ratios are 0.648 for MSHAR and 0.390 for ARFIMA. This pattern confirms that the horizon-specific interpretation of the baseline results is not driven by the use of the 5-minute realized-variance measure.

The realized-kernel results also leave unchanged the comparison between machine-learning models and the stronger econometric benchmarks. The BEST_ML row shows that the best-performing machine-learning model can improve on HAR for some stocks, especially at longer horizons and under MAE. However, these improvements are not enough to overturn the econometric ranking. Relative to the best econometric benchmark selected separately by stock, horizon, and loss function, BEST_ML almost never delivers lower average loss: the share is 0.0% at $h = 1$ and $h = 5$ for all losses, and only 2.5% at $h = 22$. In contrast, BEST_ML has significantly higher loss than the best econometric benchmark for a large fraction of stocks, ranging from 52.5% to 100.0% across horizons and loss functions. Thus, while flexible machine-learning specifications can occasionally improve on the linear HAR benchmark, they do not dominate the best econometric specifications once the comparison is made against models that incorporate regime dependence or fractional integration.

Overall, the realized-kernel robustness exercise supports the same conclusions as the baseline stock-level analysis. The evidence does not appear to be an artifact of measuring realized volatility with 5-minute realized variance. Across an alternative realized-volatility proxy, the dominant forecasting gains continue to come from parsimonious econometric structure: MSHAR is strongest at short and intermediate horizons, ARFIMA becomes important at longer horizons, and machine-learning models do not systematically improve on the best econometric benchmark.

A.4 Robustness of ML models results to number of seed ensembling

To address any concern that the neural-network results may be sensitive to the small number of random initializations used in the baseline implementation, Table A5 compares the rolling

Table A4: Cross-sectional loss ratios under rolling window using MSFE loss

Model	$h = 1$	$h = 5$	$h = 22$
Panel A: Median loss ratio relative to HAR			
HAR	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]
THAR	1.067 [1.012,1.102]	1.041 [1.021,1.070]	1.039 [1.018,1.097]
STHAR	1.065 [1.019,1.143]	1.046 [1.025,1.104]	1.050 [1.016,1.088]
MSHAR	0.620 [0.525,0.707]	0.661 [0.545,0.770]	0.648 [0.527,0.722]
ARFIMA	1.558 [1.177,2.459]	0.561 [0.548,0.639]	0.390 [0.348,0.422]
XGB	1.503 [1.203,1.703]	1.220 [1.096,1.411]	1.022 [1.012,1.067]
DNN	1.106 [1.068,1.267]	1.122 [1.074,1.222]	1.147 [1.075,1.311]
LSTM	1.459 [1.182,1.676]	1.142 [1.089,1.224]	1.041 [1.009,1.072]
LSTM-A	1.467 [1.248,1.788]	1.176 [1.101,1.275]	1.042 [1.019,1.093]
GRU-A	1.434 [1.238,1.713]	1.144 [1.103,1.252]	1.047 [1.011,1.089]
Panel B: Median loss ratio relative to best econometric benchmark			
HAR	1.613 [1.415,1.906]	1.561 [1.356,1.863]	1.580 [1.388,2.038]
THAR	1.662 [1.515,2.124]	1.622 [1.417,2.001]	1.762 [1.549,2.208]
STHAR	1.704 [1.544,2.088]	1.650 [1.406,2.052]	1.784 [1.569,2.233]
MSHAR	1.000 [1.000,1.000]	1.000 [1.000,1.000]	1.000 [1.000,1.000]
ARFIMA	2.638 [2.270,2.839]	1.038 [1.000,1.062]	1.000 [1.000,1.000]
XGB	2.295 [2.050,2.814]	2.007 [1.757,2.150]	1.695 [1.521,2.157]
DNN	1.834 [1.568,2.480]	1.806 [1.479,2.236]	2.062 [1.773,2.537]
LSTM	2.205 [2.025,2.779]	1.865 [1.623,2.195]	1.717 [1.474,2.198]
LSTM-A	2.233 [2.023,2.791]	1.926 [1.682,2.243]	1.693 [1.514,2.178]
GRU-A	2.260 [2.022,2.578]	1.809 [1.633,2.210]	1.699 [1.505,2.138]

Notes: Entries report median loss ratios across stocks; interquartile ranges are in brackets. Ratios below one indicate lower loss than the corresponding benchmark. The best econometric benchmark is selected separately for each ticker, horizon, window, and metric from HAR, THAR, STHAR, MSHAR, and ARFIMA.

five-year S&P 500 results at the one-day horizon under the baseline three-seed design and an expanded design using 40 initial seeds for the machine-learning models. The econometric forecasts are unchanged across the two designs and are included as benchmarks.

The results indicate that increasing the number of seeds does not alter the main ranking conclusions. MSHAR remains the lowest-loss model for MSFE, MAE, and QLIKE, and it remains in the 5% MCS across all three loss functions. The machine-learning losses move somewhat across seed designs, as expected for stochastic training procedures, but the changes do not lead any ML model to displace the best econometric benchmark. For example, the expanded-seed design improves the MSFE of XGB and produces small improvements in several neural-network MAE and QLIKE values, but the lowest losses continue to be delivered by MSHAR. Thus, the evidence suggests that the paper's main conclusions are not driven by the use of a small number of random initializations in the baseline ML implementation.

Table A5: Performance of ML models under small vs. large number of initial seed designs under rolling 5-year window

Stock	h	HAR	THAR	STHAR	MSHAR	ARFIMA	XGB	MLP	LSTM	LSTM-A	GRU-A
Panel A: MSFE											
	MSFE	0.116	0.123	0.135	**0.097**	0.232	0.162	0.119	0.147	0.170	0.160
No seeds= 3	MAE	0.207	0.216	0.223	**0.185**	0.318	0.234	0.219	0.220	0.228	0.224
	QLIKE	0.666	0.670	0.672	0.613	0.751	0.675	0.671	0.671	0.673	0.672
	MSFE	0.116	0.123	0.135	**0.097**	0.232	0.157	0.115	0.148	0.168	0.156
No seeds= 40	MAE	0.207	0.216	0.223	**0.185**	0.318	0.229	0.213	0.219	0.225	0.220
	QLIKE	0.666	0.670	0.672	0.613	0.751	0.667	0.664	0.664	0.666	0.665

Notes: Entries report average losses over the out-of-sample period 2020–2025 for S&P500 realized volatility for the forecast horizon $h = 1$ day. The forecasting exercise is based on the extended predictor set with EN screening specification. Rolling window starts with 2015–2019 as the initial estimation sample, while the rolling specification uses a 5-year estimation window. Underlined entries indicate 5% MCS membership based on aligned daily losses; bold-underlined entries denote the lowest loss among models in the MCS.

B Forecasting Models: models, tests, and implementation specifics

B.1 Linear models

B.1.1 Heterogeneous Autoregressive (HAR) model

The HAR model of Corsi (2009) parsimoniously approximates the long-memory dynamics of realized variance by regressing future realized volatility on rolling averages of past realized volatility over daily, weekly, and monthly horizons:

$$y_{t+h} = \alpha + \beta_1 \, \overline{y}_{t,1} + \beta_5 \, \overline{y}_{t,5} + \beta_{22} \, \overline{y}_{t,22} + \varepsilon_{t+h}, \tag{23}$$

where $\overline{y}_{t,w} = w^{-1} \sum_{j=1}^{w} y_{t-j+1}$ for $w \in \{1, 5, 22\}$. The model is estimated by OLS, and the three HAR averages serve as the forced-in regressors in the Elastic Net (EN) protocol (Section 3.5) for the extended-information set variant.

B.1.2 ARFIMA model

Realized variance and its square-root transformation y_t both exhibit long-range dependence that the HAR model approximates through its multi-scale lag structure but does not explicitly parameterize (Andersen et al., 2003; Baillie, 1996). The ARFIMA(p, d, q) model (Granger and Joyeux, 1980; Hosking, 1981) provides an explicit fractional-integration framework for this long memory. The model is

$$\Phi(L) \, (1 - L)^d \, \tilde{y}_t = \Theta(L) \, \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \tag{24}$$

where $\tilde{y}_t = y_t - \mu$, $\Phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$ and $\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$ are the AR and MA lag polynomials, and $(1 - L)^d$ is the fractional differencing operator with memory parameter $d \in (-0.49, 0.99)$. Values $d \in (0, 0.5)$ correspond to stationary long memory, while $d \in [0.5, 1)$ allows near-integrated but mean-reverting dynamics. In the extended ARFIMAX specification, EN-screened auxiliary predictors $\mathbf{x}_t$ enter before fractional differencing:

$$\Phi(L) \, (1 - L)^d \left(y_t - \mu - \boldsymbol{\beta}' \mathbf{x}_t \right) = \Theta(L) \, \varepsilon_t, \tag{25}$$

with the ARMA lags serving as the forced-in block in the EN screen.[13]

[13] We considered alternative ARFIMA specifications, including fixed choices of p and q and specifications in which the ARMA orders were selected by BIC over candidate values up to $p = q = 2$. Based on log-likelihood values, parameter estimates, AIC, BIC, and recursive forecasting diagnostics across training windows and information sets, the parsimonious ARFIMA$(0, d, 1)$ specification was sufficient to capture both long-memory persistence and short-run realized-volatility dynamics. We therefore use ARFIMA$(0, d, 1)$ throughout the reported analysis, both under the benchmark HAR information set and under the EN-screened extended predictor set.

Estimation The fractional difference filter is approximated by its truncated Maclaurin expansion, $(1-L)^d x_t \approx \sum_{k=0}^{m} \pi_k x_{t-k}$, where $\pi_0 = 1$, $\pi_k = \pi_{k-1}(k-1-d)/k$ for $k \geq 1$, and the truncation lag is set adaptively as $m = \min(800, \max(120, \lfloor T^{1/2} \rfloor))$ with T denoting the training sample size.[14] The full parameter vector $(d, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \boldsymbol{\beta}, \mu, \log \sigma^2)$ is estimated by conditional maximum likelihood (CML) via numerical optimization of the Gaussian log-likelihood evaluated from the recursively computed innovations. Stationarity and invertibility of the ARMA component are enforced by a quadratic boundary penalty on the AR and MA roots, added to the log-likelihood objective. Standard errors are obtained from the numerical Hessian of the conditional log-likelihood at the optimum.

B.2 Observable Transition Regressions

The second econometric group consists of nonlinear HAR regressions in which the forecasting relation changes with an observed transition variable. THAR allows sharp regime changes at estimated threshold values, whereas STHAR allows coefficients to change smoothly through a logistic transition function. The two models therefore represent different forms of observable state dependence. To make the comparison disciplined, both models use the same pool of transition-variable candidates and the same parsimonious division between regime-switching and common-coefficient predictors.

Throughout the nonlinear HAR specifications, the regime-switching block is restricted to the intercept and the daily HAR component,

$$\mathbf{x}_t^{\mathrm{sw}} = (1, \overline{y}_{t,1})',$$

while the weekly and monthly HAR components enter as common-coefficient regressors. In the extended-information-set, all EN-selected auxiliary predictors also enter through the common-coefficient block. Thus, the nonlinear models differ in how regimes are defined—observed thresholds, smooth observed transitions, or latent Markov states—but not in which predictors are allowed to switch across regimes.

B.2.1 Threshold HAR (THAR) model

The THAR model (Hansen , 2000; Franses and van Dijk , 2000) partitions the sample into $m+1$ mutually exclusive regimes defined by m ordered thresholds $c_1 < \cdots < c_m$ on the observed transition variable z_t. Within each regime, only the intercept and the coefficient on

[14] The truncation rule follows standard practice (Baillie , 1996) and the conservative upper bound of 800 ensures adequate approxiamtion for the range of d estimates obtained, typically in the range of 0.3-0.4.

the daily HAR component are allowed to differ. The specification is

$$y_{t+h} = \sum_{r=1}^{m+1} \mathbf{1}(z_t \in \mathcal{R}_r)\boldsymbol{\beta}_r' \mathbf{x}_t^{\text{sw}} + \boldsymbol{\delta}' \mathbf{x}_t^{\text{cm}} + \varepsilon_{t+h}, \tag{26}$$

where $\mathcal{R}_r = (c_{r-1}, c_r]$, with $c_0 = -\infty$ and $c_{m+1} = +\infty$. The vector $\mathbf{x}_t^{\text{sw}} = (1, \overline{y}_{t,1})'$ contains the regime-switching intercept and daily HAR component, while $\mathbf{x}_t^{\text{cm}}$ contains the weekly and monthly HAR components and, in the extended-information set specification, any EN-selected auxiliary predictors. Hence, the weekly, monthly, and auxiliary predictors are constrained to have common coefficients across regimes.[15]

Estimation For a given threshold variable z_t and threshold count m, the thresholds $\{c_r\}$ are estimated by sequential SSR minimization. At each sequential step, candidate threshold values are restricted to the central $[10, 90]$ percentile range of z_t, so that each tentative regime contains a minimum number of observations. When m is not fixed *a priori*, BIC is minimized over $m \in \{0, 1, \ldots, M\}$ for a specified maximum M. Conditional on the selected thresholds and threshold variable, all coefficients are estimated jointly by OLS on the indicator-interacted design matrix.

B.2.2 Smooth Transition HAR (STHAR) model

The STHAR model replaces the sharp regime changes of THAR with a continuous logistic transition function (Terasvirta , 1994; van Dijk et al. , 2002). Using the same switching and common-coefficient blocks as in the THAR specification, the model is

$$y_{t+h} = \boldsymbol{\phi}' \mathbf{x}_t^{\text{cm}} + \boldsymbol{\beta}' \mathbf{x}_t^{\text{sw}} + G(z_t; \gamma, c)\, \boldsymbol{\delta}' \mathbf{x}_t^{\text{sw}} + \varepsilon_{t+h}, \tag{27}$$

where

$$G(z_t; \gamma, c) = \left[1 + \exp\{-\gamma(z_t - c)\}\right]^{-1}, \qquad \gamma > 0, \tag{28}$$

is the logistic transition function. The speed parameter γ controls the sharpness of the transition: as $\gamma \to \infty$, the transition approaches a threshold-type change, while as $\gamma \to 0$, the model collapses to the corresponding linear HAR specification (see Granger and Terasvirta , 1993; Terasvirta , 1994; Leybourne etal. , 1998; Franses and van Dijk , 2000, for extensive dis-

[15] As a robustness check, we also estimated less parsimonious THAR, STHAR, and MSHAR specifications in which all HAR components were allowed to vary across regimes. The resulting forecasts were very similar to those from the specifications reported in the paper, but the fully switching versions often produced imprecise coefficient estimates, especially for the weekly and monthly HAR components. We therefore report the parsimonious specification in which only the intercept and daily HAR component switch across regimes, while the weekly and monthly HAR components enter with common coefficients. For comparability and parsimony, the nonlinear specifications are restricted to two regimes: in THAR this corresponds to setting $m = 1$, in STHAR to a single logistic transition function, and in MSHAR to a two-state Markov chain.

cussions). The location parameter c determines the midpoint of the transition. The effective coefficient vector on the switching block is $\boldsymbol{\beta}$ when $G = 0$ and $\boldsymbol{\beta} + \boldsymbol{\delta}$ when $G = 1$, with a smooth continuum between the two regimes. As in THAR, the switching block consists only of the intercept and daily HAR component, while the weekly, monthly, and EN-selected auxiliary predictors enter with common coefficients.

Estimation All parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \gamma, c)$ are estimated jointly by minimizing the SSR (i.e. conditional non-linear least squares as discussed in van Dijk et al., 2002). Initial values for the nonlinear parameters (γ, c) are obtained by a two-dimensional grid search: γ is searched over a geometric grid ranging from $0.01/\hat{\sigma}_z$ to $25/\hat{\sigma}_z$ (where $\hat{\sigma}_z$ is the training-sample standard deviation of z_t), and c is searched over a uniform grid spanning the central $[10, 90]$ percentile range of z_t. The grid pair achieving the lowest SSR initializes an L-BFGS-B refinement (Byrd et al., 1995). The L-BFGS-B f refinement enforces box constraints to keep γ within the grid bounds and hence, maintains numerical stability. Conditional on the optimized $(\hat{\gamma}, \hat{c})$, the linear parameters are recovered by OLS on the augmented design matrix $[\mathbf{w}_t,\ \mathbf{x}_t^{\text{sw}},\ G_t \cdot \mathbf{x}_t^{\text{sw}}]$. Heteroskedasticity-robust standard errors are computed from the linearized Jacobian of all parameters, including γ and c.

B.2.3 Transition-variable candidates and linearity screening

For both THAR and STHAR, the transition variable is selected recursively within each training window from a common pre-specified pool of RV-based candidates. In the current implementation, the active candidate set is

$$\mathcal{Z} = \{z_{rv,1}, z_{rv,2}, z_{rv,3}, z_{rv,4}, z_{rv,5}, z_{rv,1}^{(5)}, z_{rv,1}^{(22)}\},$$

where the basic candidates are lagged percentage changes in realized variance,

$$z_{rv,d,t} = \frac{RV_{t-d} - RV_{t-d-1}}{RV_{t-d-1}}, \qquad d = 1, \ldots, 5, \tag{29}$$

and $z_{rv,1}^{(5)}$ and $z_{rv,1}^{(22)}$ denote the corresponding 5-day and 22-day moving averages of $z_{rv,1,t}$. More generally, the code allows the transition-candidate set to be expanded through user-specified additional columns, but in the baseline implementation reported here the search is restricted to the RV-based candidates above.

The two nonlinear models use different linearity-screening procedures. For THAR, linearity is assessed using a formal bootstrap sup-LR test of the null of a linear HAR model against a one-threshold THAR alternative. Let SSR_0 denote the null-model sum of squared residuals and let $\text{SSR}_1(z, c)$ denote the sum of squared residuals from the one-threshold model using

candidate transition variable z and threshold value c. The test statistic is

$$\text{supLR} = \sup_{z \in \mathcal{Z},\, c} T \log\left(\frac{\text{SSR}_0}{\text{SSR}_1(z,c)}\right). \tag{30}$$

Because the threshold is unidentified under the null, the finite-sample p-value is obtained by wild bootstrap. In the empirical implementation, the bootstrap uses 499 replications and Rademacher weights (Hansen , 1996). If the null is not rejected and the linearity gate is activated, the model can be returned to the linear specification for that estimation window.

For STHAR, by contrast, the linearity screen is based on a Teräsvirta-type test for smooth-transition nonlinearity (Terasvirta , 1994). For each candidate $z \in \mathcal{Z}$, the linear benchmark is augmented by auxiliary regressors formed from the switching block interacted with z_t, z_t^2, and z_t^3. The resulting LM and F statistics are computed candidate by candidate, and the transition variable is selected according to the chosen screening rule. In the current implementation, the decision is based on the LM p-value and the selected candidate is the one with the smallest p-value. When the linearity screen is enabled, the STHAR model is then estimated either using the test-selected transition variable or, if desired, by re-selecting the transition variable according to the model-fit criterion (such as minimum SSR) within the smooth-transition specification itself. Unlike THAR, therefore, the STHAR linearity screen is not based on the bootstrap sup-LR test, but on a separate Teräsvirta-style smooth-transition test.

B.3 Latent Regime-Switching Regression

B.3.1 Markov-Switching HAR (MSHAR) model

The MSHAR model accommodates unobserved structural change in the realized-volatility process by allowing the forecasting relation to depend on a latent state governed by a Markov chain (Hamilton , 1989, 1994). Unlike THAR and STHAR, the regime at each date is not determined by an observed transition variable; instead, it is inferred from the history of volatility through filtered regime probabilities. To keep the specification comparable to the observable-transition models, the same parsimonious switching block is used: only the intercept and the daily HAR component are allowed to vary across latent states, while the weekly, monthly, and EN-selected auxiliary predictors enter with common coefficients.

The two-regime specification is

$$y_{t+h} = \boldsymbol{\beta}_{s_t}' \mathbf{x}_t^{\text{sw}} + \boldsymbol{\delta}' \mathbf{x}_t^{\text{cm}} + \varepsilon_{t+h}, \qquad \varepsilon_{t+h} \mid s_t \sim \mathcal{N}(0, \sigma_{s_t}^2), \tag{31}$$

where $s_t \in \{1, 2\}$ is the latent regime indicator, $\mathbf{x}_t^{\text{sw}} = (1, \overline{y}_{t,1})'$, and $\mathbf{x}_t^{\text{cm}}$ contains the common-coefficient regressors. The latent state follows a first-order homogeneous Markov chain with

transition probability matrix

$$P = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}, \qquad p_{ij} = \Pr(s_t = j \mid s_{t-1} = i). \tag{32}$$

The predictor vector $\mathbf{x}_t$ consists of the HAR block $(\overline{y}_{t,1}, \overline{y}_{t,5}, \overline{y}_{t,22})$ in the benchmark specification, or the corresponding semivariance components in the alternative variant, together with any EN-screened auxiliary predictors in the extended model. A flexible parameterization allows some coefficients to be common across regimes and others to be regime-specific.

Linearity test against Markov-switching alternatives Before estimating the nonlinear specification, we also implement a linearity/constancy test of the linear HAR null against Markov-switching alternatives. The procedure follows the logic of the score-type approach of Carrasco et al. (2014), which is attractive in this setting because it avoids full estimation of the Markov-switching model under the alternative hypothesis. Specifically, we first estimate the linear HAR null and then construct a null-based score process for the subset of coefficients allowed to vary across regimes. Let G_t denote the corresponding score vector under the linear null. The test aggregates lagged score cross-products over a grid of persistence parameters ρ, yielding statistics of the form

$$TS(\rho) = T \sum_{\ell=1}^{L} \left\| \rho^{\ell} \frac{1}{T} \sum_{t=\ell+1}^{T} G_t G_{t-\ell}' \right\|_F^2, \tag{33}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, so that $\|A\|_F^2 = \mathrm{tr}(A'A)$. Supremum- and exponential-type statistics are then computed over the grid of ρ values. Because the finite-sample distribution is nonstandard, p-values are obtained by a bootstrap carried out under the linear null. Rejection of the null provides evidence in favor of regime-dependent dynamics and thus supports the use of the MSHAR specification. In the empirical implementation, this test is computed recursively in each training window and recorded as a diagnostic for model selection and interpretation.

Estimation and forecasting Conditional on the selected specification, the MSHAR model is estimated by maximum likelihood. The latent-state probabilities are updated recursively using the Hamilton filter, and the fitted model delivers filtered regime probabilities for each date in the estimation window. Out-of-sample forecasts are generated by appending the test-period observations to the training sample and applying the filter with the training-window parameter estimates held fixed. The forecast for each test day is the regime-probability-weighted conditional mean,

$$\hat{y}_{t+h} = \sum_{r=1}^{2} \hat{P}(s_t = r \mid \mathcal{F}_t; \hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\beta}}_r' \mathbf{x}_t^{\mathrm{sw}} + \hat{\boldsymbol{\delta}}' \mathbf{x}_t^{\mathrm{cm}}. \tag{34}$$

where $\mathcal{F}_t$ denotes the information set available at date t and $\hat{\boldsymbol{\theta}}$ collects the parameter estimates from the training window. This forecasting design preserves the real-time nature of the exercise, since the regime probabilities used to form each prediction are conditioned only on data available up to and including the forecast date.

B.4 Gradient Boosted Trees: XGBoost

B.4.1 Model.

Gradient boosted trees (Friedman , 2001; Chen and Guestrin , 2016) produce a forecast by additively assembling K regression trees, each fitted sequentially to the negative gradient (pseudo-residuals) of the current ensemble loss:

$$\hat{y}_{t+h} = \sum_{k=1}^{K} f_k(\mathbf{x}_t), \quad f_k \in \mathcal{F}, \tag{35}$$

where $\mathcal{F}$ is the space of regression trees. Each tree is fitted using the histogram-based `hist` split algorithm with the mean squared error objective. Complexity is controlled by capping tree depth at $d_{\max}$, imposing a minimum-child-weight threshold ρ on leaf nodes, and applying an L_2 leaf-score penalty λ to the squared leaf weights. Relative to the linear econometric models, XGBoost is attractive because it can capture nonlinearities and interactions among predictors without imposing a specific parametric dynamic structure.

B.4.2 Implementation

Hyperparameters are selected by the CV procedure described above over the grid reported in Table B1. The minimum split gain $\gamma = 0$ and L_1 penalty $\alpha = 0$ are fixed at their defaults throughout. The final forecast for each test observation is the average of three independently seeded fits.

B.5 Feedforward Neural Network

B.5.1 Model

The feedforward neural network is a single-hidden-layer multilayer perceptron (MLP) that maps the standardized feature vector $\mathbf{x}_t$ to a scalar forecast through one nonlinear hidden layer:

$$\hat{y}_{t+h} = \mathbf{w}_2^{\top} \operatorname{Drop}_p\big(\sigma(\mathbf{W}_1\mathbf{x}_t + \mathbf{b}_1)\big) + b_2, \tag{36}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_u \times d_x}$ is the hidden-layer weight matrix with d_u units, $\sigma(.)$ is the logistic sigmoid activation and $\operatorname{Drop}_p(\cdot)$ is a dropout layer with rate p. The MLP serves as a flexible nonlinear benchmark that can approximate complex interactions among predictors while treating the

input at each forecast origin as a static feature vector (see, e.g., Cybenko , 1989; Goodfellow et al. , 2016).

B.5.2 Implementation

Weights are updated using the Adam optimizer (Kingma et al. , 2014) with L_2 weight decay. Overfitting is controlled by the three mechanisms described in Section B.7: a 20% validation holdout, early stopping with patience 15 and maximum 200 epochs, and dropout. The hidden-unit count d_u, learning rate η, dropout rate p, and weight decay λ_{L_2} are all selected jointly by CV.

B.6 Recurrent Neural Networks

B.6.1 LSTM

The single-layer LSTM (Hochreiter and Schmidhuber , 1997) processes each element of the input sequence through input, forget, and output gates that regulate information flow into and out of a cell state:

$$\begin{pmatrix} \mathbf{i}_\ell \\ \mathbf{f}_\ell \\ \mathbf{o}_\ell \\ \tilde{\mathbf{c}}_\ell \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(\mathbf{W} \begin{bmatrix} h_{\ell-1} \\ x_\ell \end{bmatrix} + \mathbf{b} \right), \quad \mathbf{c}_\ell = \mathbf{f}_\ell \odot \mathbf{c}_{\ell-1} + \mathbf{i}_\ell \odot \tilde{\mathbf{c}}_\ell, \quad h_\ell = \mathbf{o}_\ell \odot \tanh(\mathbf{c}_\ell), \tag{37}$$

where σ is the logistic sigmoid and $\odot$ denotes the Hadamard product. The LSTMs' main advantage is its ability to retain relevant information from earlier points in the sequence while mitigating the vanishing-gradient problem (see also Goodfellow et al. , 2016; Chung et al. , 2014). The forecast is produced by passing the final hidden state h_L through a dropout layer and a linear output head, $\hat{y}_{t+h} = \mathbf{w}^\top \mathrm{Drop}_p(h_L) + b$.

B.6.2 LSTM with Additive Attention

Compressing all temporal information into the single vector h_L may be inadequate when the relevant context is distributed across the entire window. The LSTM-Attention model augments the LSTM with an additive attention mechanism (Bahdanau et al. , 2015) that computes a learned weighted average of all L hidden states $\mathbf{H} = (h_1, \ldots, h_L)$:

$$e_\ell = \mathbf{v}^\top \tanh(\mathbf{W}_a h_\ell + \mathbf{b}_a), \quad \ell = 1, \dots, L, \tag{38}$$

$$w_\ell = \frac{\exp(e_\ell)}{\sum_{j=1}^{L} \exp(e_j)}, \qquad \mathbf{c} = \textstyle\sum_{\ell=1}^{L} w_\ell \, h_\ell, \tag{39}$$

where $\mathbf{W}_a \in \mathbb{R}^{d_a \times d_h}$ and $\mathbf{v} \in \mathbb{R}^{d_a}$ are learnable parameters. After the LSTM produces the hidden-state sequence $\{h_{t-\ell+1}, \dots, h_t\}$, the attention layer forms a context vector as a weighted average of these hidden states, with the weights determined endogenously from the data. This allows the model to emphasize the most informative parts of the recent input history when producing the volatility forecast. The context vector $\mathbf{c}$ is concatenated with the last hidden state h_L before the output head,

$$\hat{y}_{t+h} = \mathbf{w}_{\text{out}}^\top \, \text{Drop}_p\big([\mathbf{c} \parallel h_L]\big) + b_{\text{out}}, \tag{40}$$

so the output layer receives a $2d_h$-dimensional input. The attention weights $\{w_\ell\}$ are averaged across the OOS period and saved as a diagnostic on which lags in the context window receive the greatest emphasis.

B.6.3 GRU with Additive Attention

The GRU (Cho et al., 2014) is a computationally lighter alternative to the LSTM that merges the cell and hidden states and replaces the three LSTM gates with a reset gate $\mathbf{r}_\ell$ and an update gate $\mathbf{z}_\ell$:

$$\begin{aligned} \mathbf{r}_\ell &= \sigma\left(\mathbf{W}_r \begin{bmatrix} \mathbf{h}_{\ell-1} \\ \mathbf{x}_\ell \end{bmatrix}\right), & \mathbf{z}_\ell &= \sigma\left(\mathbf{W}_z \begin{bmatrix} \mathbf{h}_{\ell-1} \\ \mathbf{x}_\ell \end{bmatrix}\right), \\ \tilde{\mathbf{h}}_\ell &= \tanh\left(\mathbf{W}_h \begin{bmatrix} \mathbf{r}_\ell \odot \mathbf{h}_{\ell-1} \\ \mathbf{x}_\ell \end{bmatrix}\right), & \mathbf{h}_\ell &= (1 - \mathbf{z}_\ell) \odot \mathbf{h}_{\ell-1} + \mathbf{z}_\ell \odot \tilde{\mathbf{h}}_\ell. \end{aligned} \tag{41}$$

The GRU-Attention model applies the identical additive attention mechanism of equations (38)–(40) to the GRU hidden states. The output head, training protocol, and hyperparameter grid are identical to those of the LSTM-Attention; the attention projection dimension d_a is additionally searched over $\{16, 32\}$.

B.7 ML Models Implementation Details

B.7.1 Shared Implementation Protocol

B.7.2 Predictor construction and sequence formation.

XGBoost and the feedforward neural network use the feature vector $\mathbf{x}_t$ at each forecast origin as a static predictor vector. The recurrent models instead use the ordered sequence $\{\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t\}$, where L is the context length. Training sequences are constructed by a rolling window over the standardized training data. For each test period, the last $L-1$ observations from the training sample are prepended to the test matrix. This bridge ensures that every out-of-sample forecast is based on a complete context window and avoids a cold-start problem for the first observations in each test period.

B.7.3 Predictor supply and standardization

Each ML model is estimated under the same two information sets used for the econometric models. The benchmark variant uses only the HAR variables $(\overline{y}_{t,1}, \overline{y}_{t,5}, \overline{y}_{t,22})$. The extended variant uses the auxiliary predictors selected by the Elastic Net protocol in Section 3.5. All predictors are standardized within each recursive training window using only training-sample moments; the same means and standard deviations are then applied to the corresponding validation and test observations. This scaling rule avoids look-ahead bias and keeps the ML and econometric exercises aligned.

B.7.4 Hyperparameter tuning

Hyperparameters are selected within each training window using five-fold time-series cross-validation based on the `TimeSeriesSplit` scheme. The validation folds are ordered chronologically so that no future observation is used to tune a model for an earlier date. The tuning objective is validation MSE. XGBoost and neural-network hyperparameters are searched over the grids reported in Appendix Table B1. For the recurrent models, the sequence length L is treated as a hyperparameter and searched over $\{5, 22, 66\}$.

Table B1 summarizes the default hyperparameter configurations and cross-validation search grids for all five ML models.

B.7.5 Multi-seed ensembling

All ML specifications are estimated independently under three random seeds, $\{1, 2, 3\}$. The out-of-sample forecast is the arithmetic average of the seed-specific forecasts. For neural networks, the seed affects weight initialization and dropout masks; for XGBoost, it affects the internal random stream used in row and column subsampling. This light ensembling reduces forecast

Table B1: Default configurations and hyperparameter search grids for the ML models. All grids are searched exhaustively by 5-fold time-series CV with MSE as the tuning objective. Boldfaced values are the defaults used when tuning is disabled.

Model	Hyperparameter	Default	Search grid
XGBoost	Trees (K)	**500**	{50, 100, 300, 500, 600}
	Max depth ($d_{\max}$)	**3**	{2, 3, 4, 5}
	Learning rate (η)	**0.05**	{0.03, 0.05, 0.10}
	Row subsample (ξ_r)	**1.0**	{0.8, 1.0}
	Col subsample (ξ_c)	**1.0**	{0.8, 1.0}
	Min child weight (ρ)	**1.0**	{1.0, 5.0}
	L_2 leaf penalty (λ)	**1.0**	{1.0, 5.0}
MLP	Hidden units (d_u)	**64**	{32, 64, 128, 256}
	Learning rate (η)	$\mathbf{10^{-3}}$	$\{10^{-3},\ 5\times10^{-4}\}$
	Dropout rate (p)	**0.10**	{0.0, 0.10}
	Weight decay (λ_{L_2})	$\mathbf{10^{-4}}$	$\{10^{-5},\ 10^{-4}\}$
LSTM	Sequence length (L)	**22**	{5, 22, 66}
	Hidden size (d_h)	**64**	{32, 64, 128, 256}
	Learning rate (η)	$\mathbf{10^{-3}}$	$\{10^{-3},\ 5\times10^{-4}\}$
	Dropout rate (p)	**0.10**	{0.0, 0.10}
	Weight decay (λ_{L_2})	$\mathbf{10^{-4}}$	$\{10^{-5},\ 10^{-4}\}$
LSTM-Attention	Sequence length (L)	**22**	{5, 22, 66}
	Hidden size (d_h)	**64**	{32, 64, 128, 256}
	Attention dim (d_a)	**32**	{16, 32}
	Learning rate (η)	$\mathbf{10^{-3}}$	$\{10^{-3},\ 5\times10^{-4}\}$
	Dropout rate (p)	**0.10**	{0.0, 0.10}
	Weight decay (λ_{L_2})	$\mathbf{10^{-4}}$	$\{10^{-5},\ 10^{-4}\}$
GRU-Attention	Sequence length (L)	**22**	{5, 22, , 66}
	Hidden size (d_h)	**64**	{32, 64, 128, 256}
	Attention dim (d_a)	**32**	{16, 32}
	Learning rate (η)	$\mathbf{10^{-3}}$	$\{10^{-3},\ 5\times10^{-4}\}$
	Dropout rate (p)	**0.10**	{0.0, 0.10}
	Weight decay (λ_{L_2})	$\mathbf{10^{-4}}$	$\{10^{-5},\ 10^{-4}\}$

Notes: All neural network models use the Adam optimizer (Kingma et al. , 2014), MSE training loss, mini-batches of 128 observations in chronological order, early stopping with patience 15 (maximum 200 epochs), a 20% validation holdout (minimum 100 observations), and three-seed ensembling over seeds $\{1, 2, 3\}$.

variability arising from random initialization and stochastic training (Lakshminarayanan et al. , 2017).

B.7.6 Overfitting control for neural networks

The neural-network models are implemented in PyTorch and use several safeguards against overfitting. First, hyperparameters are selected by the time-series cross-validation procedure described above, so that model capacity is chosen using only information available within each recursive training window. Second, after the cross-validated hyperparameters are selected, the final neural-network fit uses the most recent 20% of the training window, subject to a minimum of 100 observations, as an internal validation sample for early stopping. Early stopping (Prechelt , 1998) terminates training when validation loss fails to improve for 15 consecutive epochs, with the maximum number of epochs capped at 200; before forecasting, the parameters are reset to the checkpoint with the lowest validation loss. Third, dropout (Srivastava etal. , 2014) and L_2 weight decay penalize overly complex fitted functions. Dropout rates, weight decay, learning rates, and architecture-specific dimensions are selected jointly by time-series cross-validation. Mini-batches are processed in chronological order without shuffling.

All neural-network specifications are intentionally parsimonious to control capacity in recursive financial time-series samples. This design should not be interpreted as an attempt to under-parameterize the ML benchmarks. Rather, it responds to the concern that deeper, narrower architectures may be difficult to regularize when the effective sample size is limited by strong persistence, sequence construction, validation splits, and the small number of stress episodes. The MLP uses a single hidden layer with width selected by time-series cross-validation, consistent with the universal-approximation result for sufficiently wide shallow networks (Cybenko , 1989; Hornik etal. , 1989). The LSTM, LSTM-attention, and GRU-attention models are also implemented with a single recurrent layer, so that temporal dependence is captured through recurrent state dynamics and, where applicable, the attention mechanism rather than through stacked recurrent depth. Model capacity is therefore controlled through cross-validated hidden dimensions, sequence length, dropout, learning rate, L_2 weight decay, early stopping, and multi-seed averaging, rather than through additional hidden or recurrent layers.

C Data, predictors, sample coverage, and summary statistics for equities

C.1 Data and predictors

This appendix provides additional details on the predictors used in the extended-information set specification and key relevant sample summary statistics.[16] The besaline information set contains the standard HAR components, which summarize short-, weekly-, and monthly-horizon realized-volatility persistence. The extended predictor set is designed to capture several economically distinct channels that may affect future realized volatility: return asymmetry, option-implied uncertainty, macroeconomic and monetary-policy conditions, policy uncertainty, news sentiment, systemic financial stress, geopolitical risk, broad equity-market risk factors, and, in the stock-level analysis, firm-specific trading activity.

The return-asymmetry variables are included because adverse returns are often associated with stronger subsequent volatility responses than positive returns of similar magnitude. The market-wide and macro-financial variables capture forward-looking risk appetite, business conditions, the stance and slope of the yield curve, and uncertainty about policy and geopolitical events. The Fama–French factors summarize broad equity-market risk conditions, while the stock-specific volume and trade-count variables proxy for firm-level information arrival, investor attention, and trading intensity. Table C1 summarizes the variables, definitions, economic interpretation, and data sources.

All variables are aligned to the forecast-origin date. For horizon $h \in \{1, 5, 22\}$, the target is y_{t+h}, while the HAR components and auxiliary predictors are measured using information available no later than date t. In particular, the HAR variables are computed at the forecast origin: $\bar{y}_{t,1} = y_t$, $\bar{y}_{t,5}$ is the average over the most recent five trading days ending at t, and $\bar{y}_{t,22}$ is the corresponding 22-day average ending at t. Thus, the HAR variables are contemporaneous with the forecast origin but predetermined relative to the target y_{t+h}.

C.2 Summary statistics for equities and matched S&P 500

Table C2 shows sample summary statistics for each of the 40 individual stocks plus the matched S&P500 sample (2015-2025). Average realized volatility is generally higher for individual equities than for the matched S&P 500 series, reflecting the larger idiosyncratic component of firm-level volatility. There is also meaningful cross-sectional heterogeneity: technology and high-growth stocks such as AMD, TSLA, NVDA, and NFLX exhibit among the highest av-

[16] For expositional convenience, we refer to the baseline information set as the HAR information set. Under ARFIMA, the benchmark specification is the univariate long-memory model; the ARFIMA-X specification augments that structure with the same EN-screened auxiliary predictors used in the extended information set exercises.

Table C1: Description, relevance, and sources of variables

Variable	Description	Why relevant for RV forecasting	Source
$\bar{y}_{t,1}$	1-day moving average of realized volatility	Captures very short-run volatility persistence	LSEQ/FRB/VOLARE
$\bar{y}_{t,5}$	5-day moving average of realized volatility	Captures weekly volatility persistence	LSEQ/FRB/VOLARE
$\bar{y}_{t,22}$	22-day moving average of realized volatility	Captures monthly volatility persistence and long-memory behavior	LSEQ/FRB/VOLARE
$r^-_{d,t}$	1-day moving average of negative log returns	Captures short-run asymmetric return-volatility effects	LSEQ/FRB/VOLARE
$r^-_{w,t}$	5-day moving average of negative log returns	Captures medium-run asymmetric return-volatility effects	LSEQ/FRB/VOLARE
$r^-_{m,t}$	22-day moving average of negative log returns	Captures persistent downside-risk effects	LSEQ/FRB/VOLARE
$\Delta T3M$	Daily change in 3-month Treasury yield	Proxies for monetary-policy news and front-end rate conditions	FRED
$T10Y3M$	10-year minus 3-month Treasury yield spread	Captures the slope of the yield curve and broader macro-financial conditions; ; may help proxy expectations about growth, policy, and recession risk relevant for future volatility	FRED
$\log(VIX_t)$	Log VIX index	Measures forward-looking market uncertainty and risk appetite	CBOE / FRED
ADS_t	Aruoba–Diebold–Scotti business conditions index	Tracks high-frequency changes in aggregate business conditions	Philadelphia Fed
$\log(EPU_t)$	Log Economic Policy Uncertainty index	Captures policy-related uncertainty affecting financial markets	Baker, Bloom, and Davis index
$NEWS_t$	Daily News Sentiment Index	Measures high-frequency news-based economic sentiment	Federal Reserve Bank of San Francisco
$\log(CISS_t)$	Log U.S. Composite Indicator of Systemic Stress	Captures systemic financial stress across market segments	ECB Data Portal
$\log(GPR_t^{AI})$	Log AI-GPR index	Captures geopolitical risk identified from news text using AI methods	Iacoviello and Tong website / paper
Rm_Rf_t	U.S. market excess return factor	Measures broad equity-market risk conditions	Kenneth French Data Library
SMB_t	Size factor	Captures time variation in size-related risk conditions	Kenneth French Data Library
HML_t	Value factor	Captures time variation in value-growth risk conditions	Kenneth French Data Library
RMW_t	Profitability factor	Captures profitability-related risk conditions	Kenneth French Data Library
CMA_t	Investment factor	Captures investment-related risk conditions	Kenneth French Data Library
$\Delta \log(\text{Volume}_t)$	Log daily change in trading volume (stocks only)	Proxies for firm-level information arrival and trading intensity	VOLARE
$\Delta \log(\text{Trades}_t)$	Log daily change in number of trades (stocks only)	Proxies for investor activity and market participation	VOLARE

Notes: The benchmark market RV series is based on 5-minute intraday S&P 500 data. The stock-level realized-volatility measures and stock-specific activity variables are obtained from the VOLARE database. The U.S. New CISS series is the daily U.S. systemic-stress indicator provided by the ECB. The AI-GPR variable is the daily AI-based geopolitical-risk index of Iacoviello and Tong (2026). The five Fama-French factors are the market, size, value, profitability, and investment factors from Kenneth French's Data Library. The yield-curve slope is measured as the 10-year Treasury yield minus the 3-month Treasury yield. All transformations are the authors' own calculations. All non-HAR predictors are lagged by one day.

erage realized volatilities, while more defensive names such as KO, PG, JNJ, MCD, VZ, and WMT are closer to the lower end of the distribution.

The realized-volatility series are strongly right-skewed and leptokurtic for all assets, consistent with infrequent but large volatility spikes during stress episodes. The stock-specific activity variables, measured as daily changes in log volume and log number of trades, have means close to zero but nontrivial dispersion. This pattern is consistent with these variables capturing short-run fluctuations in trading intensity and information arrival rather than persistent trends in activity levels.

Table C2: Matched-sample stock and S&P 500 summary statistics

Asset	RV mean	RV std.	RV skew.	RV ex. kurt.	$\Delta \log Vol.$ mean	$\Delta \log Vol.$ std.	$\Delta \log Trades$ mean	$\Delta \log Trades$ std.
AAPL	1.264	0.675	3.396	20.960	-0.000	0.314	0.000	0.276
ADBE	1.477	0.720	2.623	12.892	-0.000	0.358	0.000	0.291
AMD	2.702	1.072	1.633	4.452	0.000	0.360	0.001	0.332
AMGN	1.340	0.727	6.315	75.473	-0.000	0.369	0.000	0.269
AMZN	1.511	1.037	5.485	57.354	0.001	0.332	0.001	0.300
AXP	1.240	0.719	3.835	25.105	-0.000	0.374	0.000	0.284
BA	1.628	1.088	4.835	44.670	0.000	0.401	0.000	0.350
CAT	1.406	0.638	3.232	20.838	-0.000	0.372	0.000	0.296
CRM	1.566	0.827	3.836	29.792	0.000	0.390	0.000	0.322
CSCO	1.129	0.598	4.973	42.576	-0.000	0.339	0.000	0.256
CVX	1.270	0.743	4.545	33.621	-0.000	0.323	0.000	0.244
DIS	1.219	0.684	3.537	21.900	0.000	0.366	0.000	0.302
GE	1.571	0.915	2.740	13.132	-0.001	0.360	-0.000	0.293
GOOGL	1.370	0.768	3.589	23.925	0.001	0.346	0.001	0.299
GS	1.343	0.647	4.240	33.210	-0.000	0.356	0.000	0.287
HD	1.162	0.673	5.591	51.790	-0.000	0.338	0.000	0.244
HON	1.059	0.620	5.290	50.171	0.000	0.354	0.000	0.264
IBM	1.063	0.568	4.062	28.453	-0.000	0.390	0.000	0.315
JNJ	0.937	0.551	8.174	125.525	-0.000	0.337	0.000	0.245
JPM	1.189	0.656	4.646	36.500	-0.000	0.313	0.000	0.254
KO	0.880	0.492	6.013	60.165	-0.000	0.332	0.000	0.234
MCD	0.957	0.584	6.533	64.903	-0.000	0.341	0.000	0.254
META	1.563	0.779	2.259	10.504	-0.000	0.379	0.000	0.338
MMM	1.125	0.602	3.658	23.844	-0.000	0.369	0.000	0.266
MRK	1.134	0.543	3.888	27.535	0.000	0.359	0.000	0.263
MSFT	1.214	0.689	3.800	23.841	-0.000	0.315	0.000	0.262
NFLX	1.957	1.198	5.103	46.419	0.001	0.388	0.001	0.341
NKE	1.290	0.675	4.475	33.814	0.001	0.367	0.001	0.292
NVDA	2.111	1.015	2.221	8.612	0.001	0.331	0.002	0.315
ORCL	1.262	0.751	3.575	21.305	-0.000	0.357	0.000	0.283
PG	0.922	0.543	6.732	69.381	-0.000	0.335	0.000	0.227
PM	1.089	0.621	5.719	51.783	-0.000	0.362	0.000	0.267
SHW	1.302	0.681	4.649	39.360	0.000	0.374	0.001	0.275
TRV	1.094	0.581	4.479	36.869	-0.000	0.362	0.000	0.251
TSLA	2.553	1.232	2.556	11.232	0.001	0.349	0.001	0.357
UNH	1.294	0.695	4.017	28.165	0.000	0.364	0.001	0.280
V	1.113	0.652	5.331	55.977	0.000	0.358	0.001	0.251
VZ	0.998	0.496	5.091	46.761	0.000	0.331	0.000	0.260
WMT	0.998	0.545	6.368	76.730	0.000	0.348	0.000	0.271
XOM	1.248	0.735	3.948	26.516	-0.000	0.303	0.000	0.236
S&P 500	0.870	0.640	4.250	31.547				

Notes: The table reports the matched 2015–2025 sample. Realized volatility is measured as $100\sqrt{RV_t}$, where the stock realized-variance measure is the RV5 in the VOLARE dataset. The two stock-specific predictors are daily changes in log volume and log number of trades.

C.3 Sample coverage and forecast counts

Table C3 documents the effective estimation-sample sizes and out-of-sample forecast counts used in the recursive forecasting exercise. Because all 40 individual equities share the same matched sample dates and observation-count structure, the stock-level counts are reported once rather than repeated separately for each ticker. The table also reports the corresponding counts for the long-sample S&P 500 exercise to clarify differences in sample length and rolling-window construction across the two applications. As explained in Section 3.3, the out-of-sample evaluation sample is indexed by forecast-origin dates. For each trading day t in a test year, the models use information available at the forecast origin to predict y_{t+h}, and the forecast remains assigned to the year containing t even when the target date $t+h$ falls in the following calendar year. The effective estimation sample is horizon specific: at each re-estimation date, a training observation is included only if its associated target y_{t+h} has already been realized, so that no future target information enters model estimation. Accordingly, $N_{\text{train}}^{(h)}$ generally declines with the forecast horizon. The number of valid out-of-sample forecasts, $N_{\text{OOS}}^{(h)}$, is likewise reported separately by horizon. These counts need not equal the number of observations in a calendar year minus h and may differ across horizons because of target availability and complete-case alignment of the predictor set, particularly near the end of the sample or when missing observations propagate through rolling predictors.

Table C3: Effective training-sample sizes and out-of-sample forecast counts

Test year	Expanding $N_{\text{train}}^{(1/5/22)}$	Rolling $N_{\text{train}}^{(1/5/22)}$	$N_{\text{OOS}}^{(1/5/22)}$
Panel A: 40 individual stocks, matched sample (expanding and rolling 5-year window)			
2020	1,234 / 1,230 / 1,213	1,234 / 1,230 / 1,213	253 / 253 / 253
2021	1,487 / 1,483 / 1,466	1,258 / 1,254 / 1,237	252 / 252 / 252
2022	1,739 / 1,735 / 1,718	1,258 / 1,254 / 1,237	251 / 251 / 251
2023	1,990 / 1,986 / 1,969	1,258 / 1,254 / 1,237	250 / 250 / 250
2024	2,240 / 2,236 / 2,219	1,257 / 1,253 / 1,236	250 / 250 / 250
2025	2,490 / 2,486 / 2,469	1,255 / 1,251 / 1,234	224 / 220 / 203
Total OOS	–	–	1,480 / 1,476 / 1,459
Panel B: S&P 500, long sample (expanding and rolling 10-year window)			
2006	2,416 / 2,412 / 2,395	2,416 / 2,412 / 2,395	251 / 251 / 251
2007	2,667 / 2,663 / 2,646	2,468 / 2,464 / 2,447	251 / 251 / 251
2008	2,918 / 2,914 / 2,897	2,472 / 2,468 / 2,451	253 / 253 / 253
2009	3,171 / 3,167 / 3,150	2,480 / 2,476 / 2,459	252 / 252 / 252
2010	3,423 / 3,419 / 3,402	2,482 / 2,478 / 2,461	252 / 252 / 252
2011	3,675 / 3,671 / 3,654	2,483 / 2,479 / 2,462	252 / 252 / 252
2012	3,927 / 3,923 / 3,906	2,512 / 2,508 / 2,491	250 / 250 / 250
2013	4,177 / 4,173 / 4,156	2,511 / 2,507 / 2,490	252 / 252 / 252
2014	4,429 / 4,425 / 4,408	2,513 / 2,509 / 2,492	252 / 252 / 252
2015	4,681 / 4,677 / 4,660	2,516 / 2,512 / 2,495	252 / 252 / 252
2016	4,933 / 4,929 / 4,912	2,516 / 2,512 / 2,495	252 / 252 / 252
2017	5,185 / 5,181 / 5,164	2,517 / 2,513 / 2,496	251 / 251 / 251
2018	5,436 / 5,432 / 5,415	2,517 / 2,513 / 2,496	251 / 251 / 251
2019	5,687 / 5,683 / 5,666	2,515 / 2,511 / 2,494	247 / 247 / 246
2020	5,935 / 5,935 / 5,918	2,511 / 2,511 / 2,494	230 / 230 / 230
2021	6,164 / 6,160 / 6,142	2,488 / 2,484 / 2,466	252 / 252 / 252
2022	6,416 / 6,412 / 6,394	2,488 / 2,484 / 2,466	251 / 251 / 251
2023	6,667 / 6,663 / 6,645	2,489 / 2,485 / 2,467	236 / 236 / 236
2024	6,904 / 6,900 / 6,882	2,474 / 2,470 / 2,452	249 / 249 / 249
2025	7,153 / 7,149 / 7,131	2,471 / 2,468 / 2,450	224 / 220 / 203
Total OOS	–	–	4,960 / 4,955 / 4,937

Notes: Entries separated by slashes are reported in the order $h = 1/5/22$. $N_{\text{train}}^{(h)}$ is the effective number of observations used to estimate the horizon-h model at the beginning of the corresponding test period. The effective training sample is horizon specific: an origin-dated observation is retained for estimation only when its associated target y_{t+h} has been realized before the new out-of-sample period begins. Thus, the reported training counts incorporate the horizon-specific target-availability embargo as well as the complete-case alignment of the forecasting dataset. $N_{\text{OOS}}^{(h)}$ is the number of valid daily forecast-origin/target pairs entering evaluation at horizon h. Evaluation years are indexed by forecast-origin date t, not by target date $t+h$. Accordingly, a forecast formed in year Y remains assigned to year Y even when its realization falls in the following calendar year. The OOS counts are therefore not constructed as non-overlapping h-day blocks and are not mechanically equal to the number of observations in a calendar year minus h. Counts can nevertheless differ across horizons when complete-case alignment or future-target availability differs, particularly near the end of the sample. Panel A reports the common matched-stock design using the AAPL count file as the representative stock-level record. The stock forecasting exercise is evaluated over 2020–2025 under expanding and rolling five-year windows. Panel B reports the long-sample S&P 500 exercise evaluated over 2006–2025 under expanding and rolling ten-year windows. The "Total OOS" row sums valid forecasts over the reported test years; training counts are not summed because the recursive estimation windows overlap across test periods.

D Feature importance in ML models and parameter estimates in econometric models

D.1 Feature importance and predictor selection in machine-learning models

To assess model explainability, this appendix summarizes how the machine-learning models allocate importance across predictors under the rolling-window, yearly-retraining extended-information set specification. For the neural-network models (MLP, LSTM, LSTM-A, and GRU-A), feature importance is measured using mean SHAP importance aggregated across re-estimation windows. For XGBoost, feature importance is measured using the gain-based built-in importance metric, again aggregated across re-estimation windows. SHAP provides an additive decomposition of a prediction into feature-level contributions and is therefore especially useful for model explainability at both the local and global levels (Lundberg and Lee , 2017). By contrast, XGBoost's built-in importance is an internal tree-based diagnostic. Under the default `gain` definition used by `feature_importances_`, a feature is important to the extent that splits on that feature produce large average reductions in the objective function (XGBoost Developers , 2025). Accordingly, the SHAP-based measures for the neural-network models and the built-in XGBoost importance are not numerically identical objects, but both are informative summaries of relative predictor relevance. In addition, we report predictor-inclusion frequencies and average feature-rank heatmaps in order to summarize how often a predictor enters the final training set and how highly it ranks within each model-horizon combination. The discussion below focuses on the yearly rolling-window results reported in Figures D1–D4; qualitatively similar patterns are obtained under the expanding-window specification, and those additional figures are available upon request.

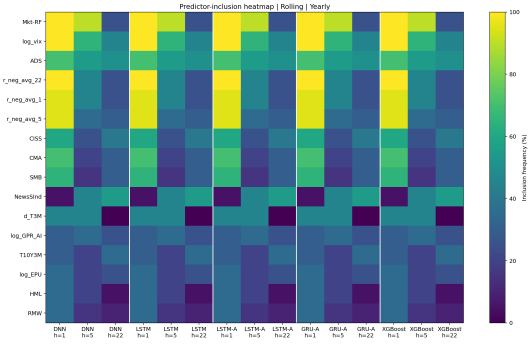
The main message is that the machine-learning models do not behave as black boxes in the sense of relying on unstable or economically uninterpretable predictors. Instead, the same broad set of variables repeatedly appears among the most important features across models and horizons. In particular, log(VIX) is consistently one of the most important predictors across all machine-learning architectures and forecast horizons; this is evident both from the horizon-specific importance profiles in Figures D2–D4 and from the summary heatmaps in Figure D1. The HAR persistence terms are also central, with $\bar{y}_{t,1}$ and $\bar{y}_{t,5}$ especially prominent at short and medium horizons, and $\bar{y}_{t,22}$ becoming relatively more important at the long horizon. At the same time, the importance of broader macro-financial and sentiment variables rises with the forecast horizon: NewsSInd, ADS, $T10Y3M$, CISS, and several factor variables tend to rank more highly at $h = 22$ than at $h = 1$; compare, in particular, Figures D2 and D4.

A second regularity is that short-horizon forecasts place comparatively greater weight on asymmetric-return information. The negative-return aggregates, especially $r\_neg\_avg\_1$, $r\_neg\_avg\_5$, and $r\_neg\_avg\_22$, are selected frequently and often rank highly at $h = 1$ and, to a lesser ex-

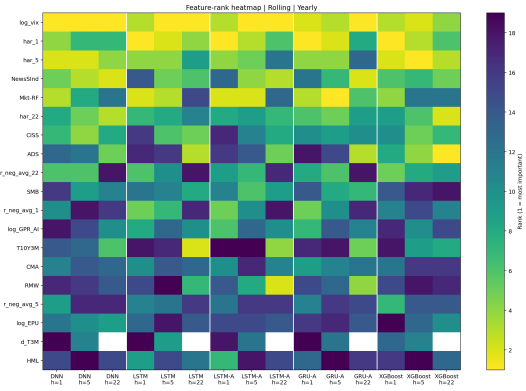
tent, at $h = 5$; see Figures D2 and D3. By contrast, their importance generally declines at the 22-day horizon, where the models place relatively more weight on longer-run HAR persistence and macro-financial conditions. The term-structure and sentiment variables exhibit the opposite pattern: they are present at shorter horizons but become more central at the medium and long horizons, as can be seen by the stronger role of NewsSInd, ADS, $T10Y3M$, and CISS in Figure D4.

These patterns are visible both in the horizon-specific bar charts and in the two heatmaps. The predictor-inclusion heatmap in Figure D1 shows that a relatively stable subset of predictors—especially log(VIX), HAR terms, and several macro-financial variables—enters the final training sets repeatedly across yearly rolling retrainings. The feature-rank heatmap in Figure D1 shows that the ordering of predictors changes systematically with the forecast horizon rather than randomly across architectures. Taken together, Figures D1–D4 indicate that the machine-learning models learn economically sensible relationships: near-term forecasts depend primarily on short-run volatility persistence and current market conditions, whereas medium- and long-horizon forecasts place greater weight on broader macro-financial and sentiment indicators. As noted above, the corresponding figures under the expanding-window specification deliver the same qualitative message and are available upon request.

Figure D1: Summary explainability diagnostics for machine-learning models under the rolling-window, yearly-retraining extended-information set specification.

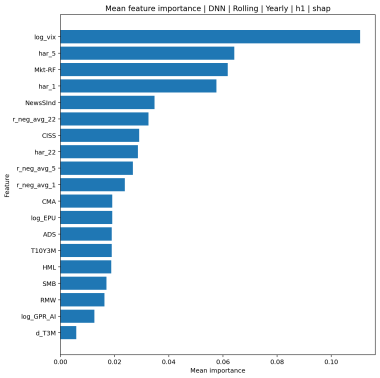


(a) Predictor-inclusion frequencies across models and horizons.

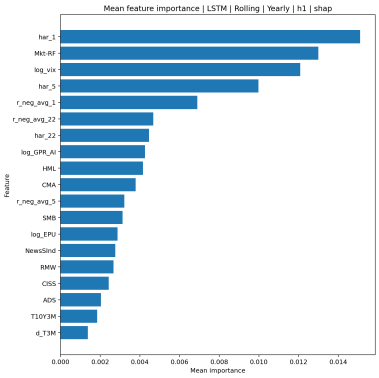


(b) Average feature rank across models and horizons.

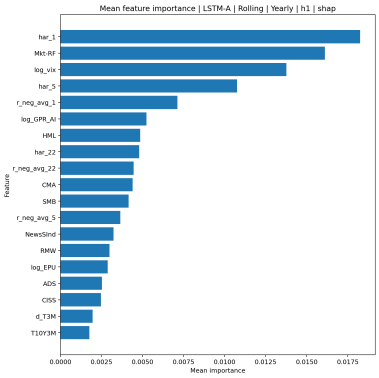
Notes: The left panel reports the fraction of re-estimation windows in which a predictor is included in the final model. The right panel reports the average rank of each predictor within a given model-horizon combination, where rank 1 denotes the most important predictor. For the neural-network models, ranks are based on mean SHAP importance; for XGBoost, they are based on the built-in feature-importance measure.



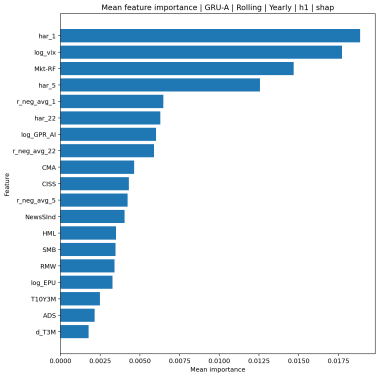
(a) MLP



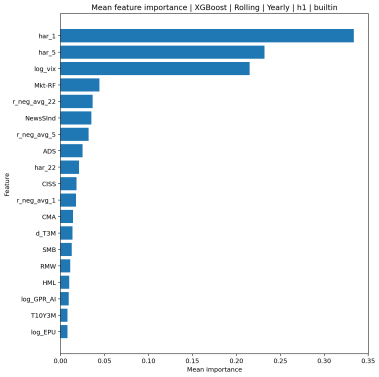
(b) LSTM



(c) LSTM-A

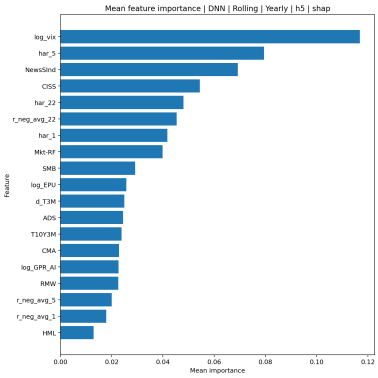


(d) GRU-A

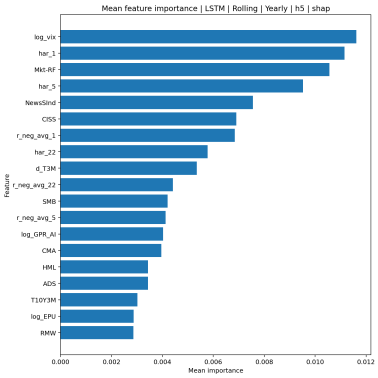


(e) XGBoost

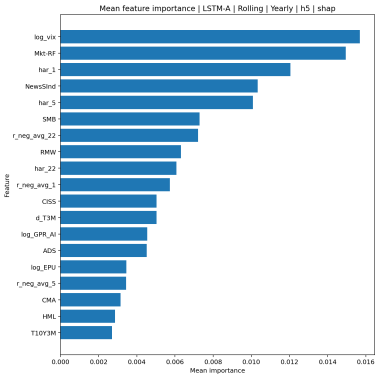
Figure D2: Mean feature importance at the 1-day horizon under the rolling-window, yearly-retraining extended-information set specification. The neural-network bars are based on mean SHAP importance, while the XGBoost bars use the model's built-in feature-importance measure. The short-horizon figures show a strong role for log(VIX), the short-run HAR terms, and asymmetric-return predictors.



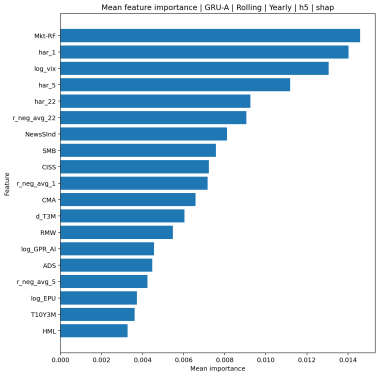
(a) MLP



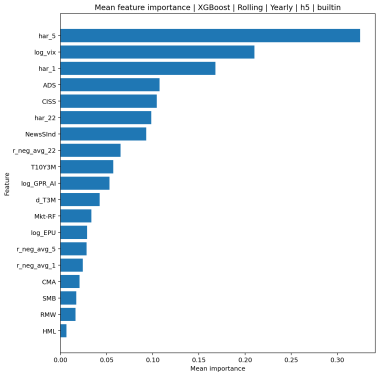
(b) LSTM



(c) LSTM-A

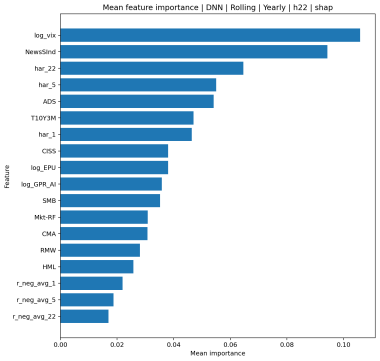


(d) GRU-A

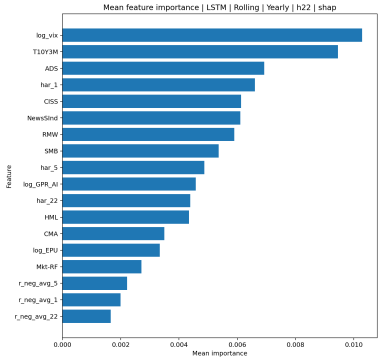


(e) XGBoost

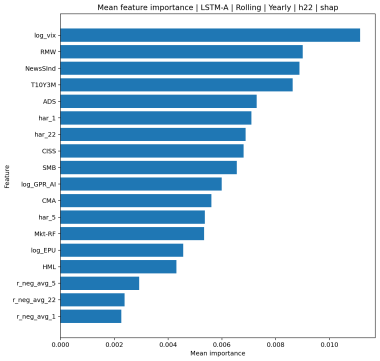
Figure D3: Mean feature importance at the 5-day horizon under the rolling-window, yearly-retraining extended-information set specification. Relative to the 1-day horizon, short-run HAR persistence remains important, but broader variables such as NewsSInd, ADS, CISS, and selected macro-financial indicators become more prominent.



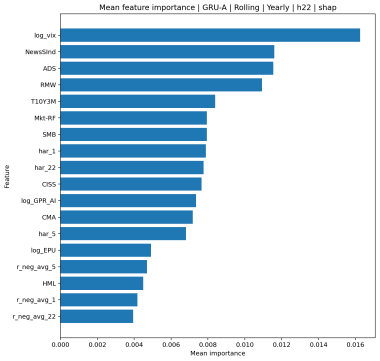
(a) MLP



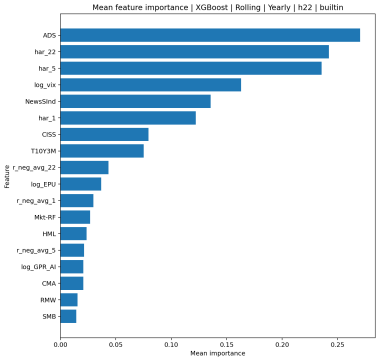
(b) LSTM



(c) LSTM-A



(d) GRU-A



(e) XGBoost

Figure D4: Mean feature importance at the 22-day horizon under the rolling-window, yearly-retraining extended information-set specification. At the long horizon, the importance profile shifts further toward medium-run HAR persistence and broader macro-financial and sentiment variables such as NewsSInd, ADS, $T10Y3M$, CISS, and factor variables, although log(VIX) remains important across all models.

D.2 Parameter estimates for econometric models

To provide additional transparency on the estimated econometric specifications, Table D1 reports a representative set of parameter estimates for the rolling 10-year window, one-day-ahead forecasts, and the 2025 test year. The table summarizes the HAR, THAR, STHAR, MSHAR, and ARFIMA estimates using a common notation for the intercept and HAR coefficients, while also reporting model-specific parameters such as threshold, transition, Markov-transition, and long-memory parameters. The estimates are intended to illustrate the typical parameter magnitudes and the way regime dependence and long memory enter the forecasting models. Complete yearly parameter estimates for all horizons, windows, and test periods are available upon request.

Table D1: Parameter estimates for econometric models, rolling 10-year window, $h = 1$, test year 2025

Parameter	HAR	THAR	STHAR	MSHAR	ARFIMA
β_0	0.067***	-0.015	-0.014	0.093***	
	(0.019)	(0.016)	(0.028)	(0.009)	
β_0^*		0.117***	0.157**	0.485***	
		(0.013)	(0.065)	(0.066)	
β_1	0.507***	0.792***	0.978***	0.426***	
	(0.055)	(0.037)	(0.160)	(0.019)	
β_1^*		0.527***	-0.428***	0.928***	
		(0.025)	(0.160)	(0.043)	
β_5	0.377***	0.272***	0.170***	0.342***	
	(0.069)	(0.035)	(0.043)	(0.027)	
β_{22}	0.019	0.019	0.032	0.051**	
	(0.046)	(0.024)	(0.025)	(0.021)	
μ					0.682***
					(0.038)
d					0.301***
					(0.011)
θ					0.206***
					(0.007)
c		-0.207	-0.365		
			(0.247)		
γ			2.975**		
			(1.210)		
p_{00}				0.973***	
				(0.005)	
p_{10}				0.840***	
				(0.048)	
$p_{11} = 1 - p_{10}$				0.160	
σ^2				0.044***	
				(0.001)	0.089
Selected z_t		z_rv_1	z_rv_1		

Notes: Entries are estimated using the rolling 10-year training sample ending in 2024 and are associated with the 2025 out-of-sample forecast period at horizon $h = 1$. HAC robust standard errors are reported in parentheses beneath coefficient estimates. *, **, and *** denote significance at the 10%, 5%, and 1% levels. The notation β_0, β_1, β_5, and β_{22} refers to the intercept and the coefficients on $\bar{y}_{t,1}$, $\bar{y}_{t,5}$, and $\bar{y}_{t,22}$, respectively. For THAR and MSHAR, starred coefficients such as β_0^* and β_1^* denote second-regime coefficients. For STHAR, β_0^* and β_1^* denote smooth-transition increments to the intercept and daily HAR coefficient, so the limiting upper-regime coefficients are $\beta_0 + \beta_0^*$ and $\beta_1 + \beta_1^*$. For MSHAR, p_{00} and p_{10} are the reported transition probabilities; p_{11} is computed as $1 - p_{10}$.

E Training/validation, early stopping, dropout diagnostics for neural network models

To address the concern that the neural-network models may be prone to overfitting, we implemented a common set of safeguards across the MLP, LSTM, LSTM-A, and GRU-A specifications. In each case, hyperparameter tuning is conducted by time-series cross-validation, estimation uses dropout and L_2 regularization, and training is monitored by validation loss with early stopping; when validation loss ceases to improve for a fixed patience window, estimation stops and the model parameters are reset to the best validation-loss checkpoint rather than the final epoch. To document the realized training behavior, we compile diagnostics by forecast origin and seed, including whether early stopping was triggered, the best stopping epoch, the total number of epochs run, and the gap between the final and best validation loss. These diagnostics are informative because pervasive overfitting would be expected to show up as persistently late stopping, systematically widening train–validation gaps, and materially worse validation behavior near the end of training.

Tables E1 and E2 summarize these diagnostics under extended-information set (EN screened auxilary predictors). Several features are noteworthy. First, early stopping is triggered in nearly all runs: under yearly retraining, the early-stopping rate is typically between 95 and 100 percent across models, windows, and horizons, and the quarterly rolling results show the same pattern. Second, the median best epoch is generally well before the maximum number of epochs run, which indicates that the stopping rule is active rather than simply allowing estimation to continue to the epoch cap. Third, the mean deterioration in validation loss between the final and best checkpoints, $\overline{\Delta}_{val}$, remains small throughout, ranging from 0.0026 to 0.0380 under yearly retraining and from 0.0053 to 0.0243 under quarterly retraining. These magnitudes imply that, even when training extends somewhat beyond the validation optimum, the resulting deterioration in validation fit is modest rather than large. Results under HAR-only predictors are qualitatively similar and are available upon request.

The loss-curve evidence in Figures E1 and E2 complements the tabular diagnostics. These figures plot the mean training and validation losses across seeds for a representative case—yearly retraining, $h = 1$, and test year 2025 (i.e., training/validation ending in 2024)—with one panel for each neural-network architecture under the expanding and rolling windows, respectively. The selected examples are representative: the qualitative behavior of the curves is similar across other forecast origins and horizons. Across all eight panels, both training and validation losses decline sharply during the early epochs, and the validation loss either flattens out or rises only mildly once it reaches its minimum. Importantly, none of the panels displays the type of pronounced and persistent divergence between training and validation loss that would be expected under severe overfitting. Instead, the train–validation gap remains limited, and the

Table E1: Summary training/validation and early stopping diagnostics for neural network models under yearly retraining

Model	Win.	Hor.	ES%	Best ep.	Run ep.	$\overline{\Delta}_{val}$
MLP	Expanding	h1	98.3	54.5	69.5	0.0042
		h5	100.0	42.0	57.0	0.0026
		h22	100.0	20.0	35.0	0.0121
	Rolling	h1	100.0	58.5	73.5	0.0078
		h5	95.0	48.5	63.5	0.0037
		h22	100.0	19.0	34.0	0.0117
GRU-A	Expanding	h1	100.0	23.0	38.0	0.0139
		h5	100.0	12.5	27.5	0.0143
		h22	100.0	3.0	18.0	0.0300
	Rolling	h1	100.0	38.5	53.5	0.0091
		h5	100.0	9.5	24.5	0.0143
		h22	100.0	6.0	21.0	0.0132
LSTM	Expanding	h1	100.0	28.5	43.5	0.0126
		h5	100.0	14.0	29.0	0.0238
		h22	100.0	5.0	20.0	0.0380
	Rolling	h1	100.0	44.0	59.0	0.0111
		h5	100.0	7.5	22.5	0.0251
		h22	100.0	5.0	20.0	0.0226
LSTM-A	Expanding	h1	100.0	28.0	43.0	0.0134
		h5	100.0	14.0	29.0	0.0306
		h22	100.0	4.0	19.0	0.0263
	Rolling	h1	100.0	41.5	56.5	0.0171
		h5	96.7	8.0	23.0	0.0213
		h22	100.0	5.0	20.0	0.0232

Notes: Each row summarizes diagnostics aggregated across forecast origins and random seeds for the indicated model, window setting, and forecast horizon. "% early stopped" is the share of runs for which estimation terminated before the maximum epoch limit. "Median best epoch" is the median epoch at which the lowest validation loss was attained, while "Median epochs run" is the median total number of epochs executed. The quantity $\overline{\Delta}_{val}$ denotes the mean of final validation loss minus the best validation loss, so values near zero indicate that training stopped close to the validation optimum, whereas larger positive values indicate that continuing to the final epoch would have worsened validation performance, consistent with the logic of early stopping.

Table E2: Summary training, validation and early stopping diagnostics for neural network models under quarterly retraining)

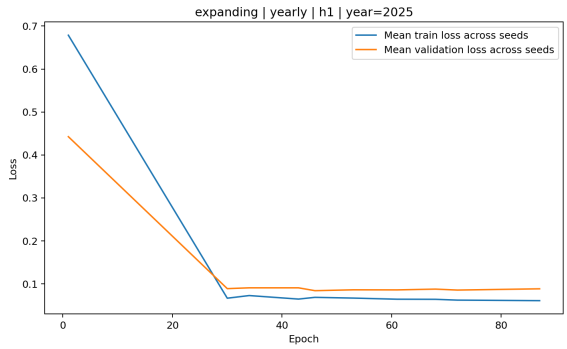
Model	Win.	Hor.	ES%	Best ep.	Run ep.	$\overline{\Delta}_{val}$
MLP	Rolling	h1	100.0	44.0	59.0	0.0053
		h5	100.0	38.0	53.0	0.0061
		h22	98.3	20.0	35.0	0.0166
GRU-A	Rolling	h1	100.0	38.5	53.5	0.0101
		h5	100.0	10.0	25.0	0.0136
		h22	100.0	6.0	21.0	0.0149
LSTM	Rolling	h1	100.0	34.0	49.0	0.0115
		h5	100.0	10.0	25.0	0.0243
		h22	100.0	6.0	21.0	0.0228
LSTM-A	Rolling	h1	100.0	41.0	56.0	0.0113
		h5	100.0	10.0	25.0	0.0198
		h22	100.0	7.0	22.0	0.0216

Notes: See, notes for Table E1.

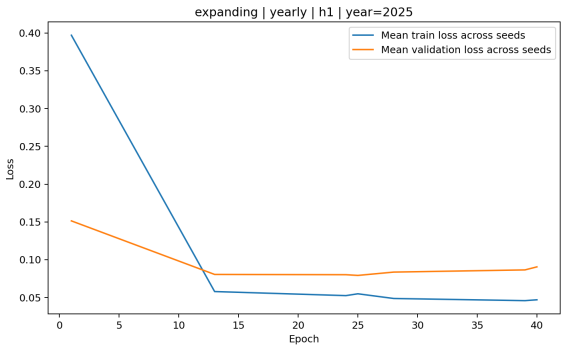
mild late-epoch increases in validation loss are precisely the cases that the early-stopping rule is designed to handle by restoring the best checkpoint rather than retaining the final one.

The graphical evidence is especially informative for understanding why the neural-network results should not be interpreted as artifacts of uncontrolled overfitting. In the rolling-window examples, the training and validation curves become very close once losses have stabilized, indicating that the models are not continuing to fit the training sample at the expense of sharply worse validation performance. In the expanding-window examples, the train–validation gap is somewhat larger in some panels, but the validation loss still remains stable and the post-minimum deterioration is gradual rather than abrupt. Taken together with the high incidence of early stopping and the small values of $\overline{\Delta}_{val}$, these patterns indicate that the implemented regularization and stopping rules are functioning as intended. Accordingly, the relatively weaker out-of-sample performance of the neural-network models cannot be attributed simply to uncontrolled overfitting in training; instead, it appears to reflect genuine differences in predictive performance relative to the leading econometric specifications.

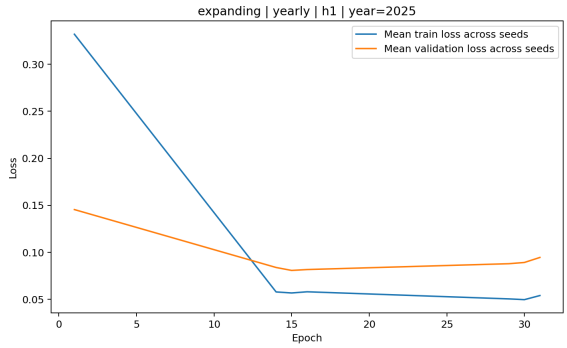
Figure E1: Mean training and validation loss across seeds under the expanding window, yearly retraining, $h = 1$, test year 2025.



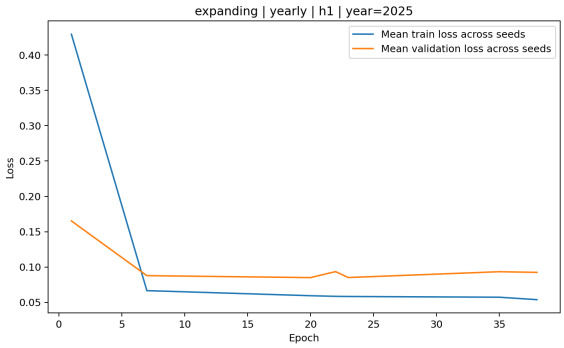
(a) MLP



(b) LSTM

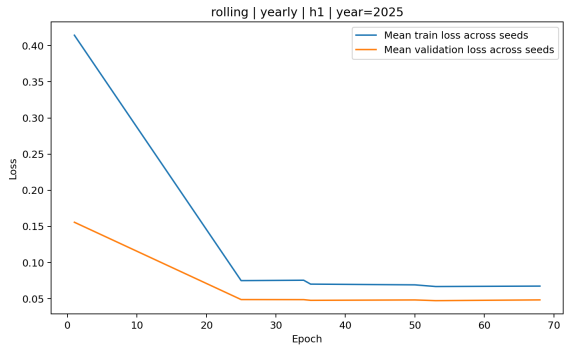


(c) LSTM-A

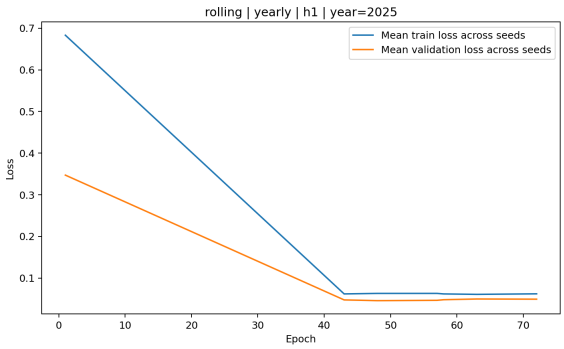


(d) GRU-A

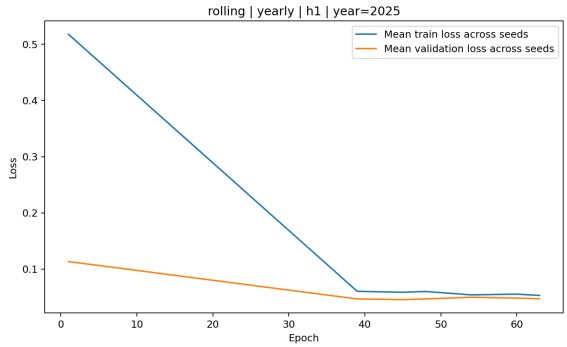
Figure E2: Mean training and validation loss across seeds under the rolling window, yearly retraining, $h = 1$, test year 2025.



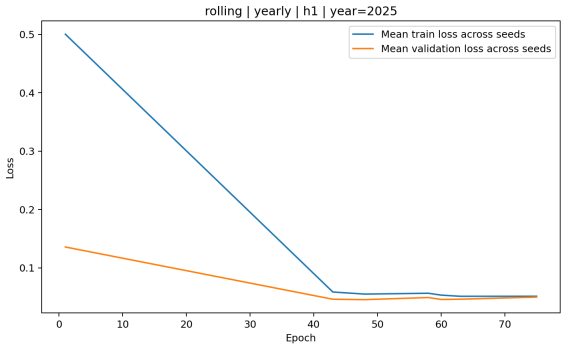
(a) MLP



(b) LSTM



(c) LSTM-A



(d) GRU-A