UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 12

Forecasting Realized Volatility Across Asset Classes

HAR, GARCH, and Machine Learning in a Unified Multi-Asset Horse Race, 2013–2026

Simon-Pierre Boucher

Département des sciences administratives
Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

October 2026

Version 0.2

Abstract

We compare 27 forecasts of daily realized variance—six HAR specifications, GARCH, GJR, EGARCH and Realized GARCH, penalized regressions, random forests, XGBoost and LightGBM on two information sets, an LSTM, a Transformer, a LightGBM pooled across assets, and two combinations—on 25 liquid instruments from four asset classes (equities and ETFs, FX, cryptocurrencies, futures) over 2013–2026, under one protocol: the same subsampled 5-minute RV target, a 1,000-day rolling window, horizons of 1, 5 and 22 days, QLIKE and MSE losses, Diebold–Mariano tests, model confidence sets, and a common evaluation sample of 59,323 asset-days. Log-linear models dominate: log-HAR lowers the QLIKE of HAR in levels by 25% at one day on every instrument, and ridge regression on an extended predictor set is the best individual forecaster (ratio 0.70), significantly better than HAR on 24 of 25 instruments. Tree ensembles and deep learners trail the linear learners and fall behind HAR at one month; pooling across assets helps the trees but does not transfer to unseen assets outside a homogeneous class. Gains are largest for cryptocurrencies (0.62) and smallest for FX (0.85), grow with the noise of the realized measure, and shrink in the calm 2024–2026 period. In volatility timing net of 5 basis points, smooth forecasters (LSTM, GARCH) earn higher Sharpe ratios and positive utility gains over HAR than the most accurate ones, which trade too much; at the 5% VaR level all forecasters are calibrated, at 1% none is.

Keywords: realized volatility, HAR model, GARCH, machine learning, forecast evaluation, model confidence set, volatility timing

JEL Classification: C22, C45, C53, G17

1. Introduction

Volatility forecasts are an input to almost every decision in financial risk management, derivatives pricing and portfolio construction, and the availability of intraday data has turned the forecasting problem into a well-posed time-series exercise: realized variance computed from high-frequency returns is a consistent, nearly observable measure of the latent quadratic variation of prices (Andersen and Bollerslev, 1998; Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2002). For two decades the heterogeneous autoregressive model of Corsi (2009) has been the benchmark of choice for forecasting realized variance: three OLS coefficients on daily, weekly and monthly averages capture most of the long-memory dynamics of volatility, and its many extensions—jump components, signed semivariances, measurement-error corrections—have produced incremental gains that are modest in size but remarkably stable across markets (Andersen et al., 2007; Patton and Sheppard, 2015; Bollerslev et al., 2016).

More recently a growing literature has asked whether statistical learning can do better. Regularized regressions, tree ensembles and neural networks have been reported to improve on HAR by ten to twenty percent in QLIKE on US equities, with the gains concentrated in turbulent periods and at longer horizons (Christensen et al., 2023; Zhang et al., 2024), while other careful comparisons on comparable data find no reliable advantage of nonlinear learners over well-specified linear models (Branco et al., 2024). The disagreement is not surprising. The published horse races differ in the realized measure used as target and as predictor, in the length and type of the estimation window, in the loss function, in whether inference accounts for the multiplicity of models compared, and, above all, in the asset universe, which is almost always a single class—usually US stocks or a single equity index. Whether any machine-learning gain survives in foreign exchange, in futures or in cryptocurrencies, whose volatility levels, trading sessions, jump intensities and microstructure noise differ by an order of magnitude from equities, is an open question, and so is the closely related question of whether a model trained on many assets can forecast an asset it has never seen.

This paper runs a single, fully controlled horse race across asset classes. We compute identical daily realized measures from 1-minute bars for 25 liquid instruments—ten equities and ETFs, five currency pairs, five cryptocurrencies and five continuous futures—over January 2013 to July 2026, and forecast the average realized variance over the next one, five and twenty-two trading days with the same rolling protocol for every model: a 1,000-day estimation window, re-estimation every 22 days, forecasts every day, and a common evaluation sample per asset. The model set spans the econometric benchmarks that a sceptical reader would want to see beaten—six HAR specifications including HARQ and the semivariance HAR, GARCH, GJR, EGARCH and the Realized GARCH—and the main families of statistical learners: ridge, LASSO and elastic net, random forests, XGBoost and LightGBM, each on the HAR information set and on an extended

set of eighteen predictors, an LSTM and a Transformer encoder, a LightGBM pooled across all 25 assets with asset identifiers, and two forecast combinations. Forecasts are evaluated with the QLIKE and MSE losses, which are robust to the noise in the realized proxy (Patton, 2011), compared pairwise with Diebold–Mariano tests and jointly with the model confidence set of Hansen et al. (2011), checked for calibration and encompassing, and tracked over time with the fluctuation test of Giacomini and Rossi (2010). The statistical comparison is then pushed through to economic value in the volatility-timing framework of Fleming et al. (2001, 2003) and to Value-at-Risk coverage.

Three findings emerge. First, the gains over HAR that the machine-learning literature reports do generalize across asset classes, but most of the gain is not machine learning: forecasting the *logarithm* of realized variance with the HAR predictors and a lognormal retransformation lowers the QLIKE of the HAR benchmark in levels by 25 percent at one day and 37 percent at one week, on all 25 instruments, and penalized linear regressions on an extended predictor set add about five points more and are the best individual forecasters, significantly better than HAR on 24 of 25 instruments and never significantly worse. Tree ensembles and deep learners, estimated asset by asset on a thousand days, are worse than the linear learners everywhere and worse than HAR at one month; pooling across assets rescues the trees but does not make them better than log-HAR, and the information they use is, by the SHAP attribution, almost entirely the HAR cascade. Second, the gains are largest where volatility is most volatile—a QLIKE ratio of 0.62 for the cryptocurrencies against 0.85 for FX—they grow with the noise of the realized measure and shrink in the calm 2024–2026 period, and a pooled model transfers to an unseen asset only inside a homogeneous class such as the currency pairs. Third, statistical accuracy and economic value diverge once transaction costs are charged: in volatility timing at 5 basis points the smoothest forecasters, the LSTM and the daily GARCH, earn the highest net Sharpe ratios and positive utility gains over HAR, whereas the most accurate forecasters trade three to nine times as much and lose 47 to 121 basis points a year relative to HAR; at the 5 percent Value-at-Risk level every forecaster is calibrated and the linear learners have the best coverage, while at 1 percent all fail because of the Gaussian tail.

The contribution of the paper is therefore threefold. First, it provides the first comparison of HAR-type, GARCH-type and machine-learning forecasts of realized volatility that holds the target, the protocol, the losses and the inference fixed across four asset classes, so that differences in model rankings across classes can be attributed to the data rather than to the design. Second, it asks directly whether volatility dynamics are transferable across assets, by training a pooled learner with and without the target asset in the training pool. Third, it reports the economic counterpart of every statistical gain within the same design. Everything from the raw bars to the tables is produced by a versioned pipeline, and every number in the paper is traceable to a machine-generated output.

The remainder of the paper proceeds as follows. Section 2 reviews the literature on realized measures, HAR and GARCH models, machine learning for volatility and forecast evaluation.

Section 3 describes the data and the realized measures. Section 4 sets out the models, the rolling protocol, the loss functions, the tests and the economic-value exercises. Section 5 presents the statistical results, Section 6 the robustness analyses, Section 7 the economic value, and Section 8 concludes.

2. Related literature

2.1. Realized volatility and its measurement

The modern volatility-forecasting literature begins with the insight that intraday returns turn latent variance into an effectively observable quantity. Andersen and Bollerslev (1998) showed that standard models forecast well once evaluated against realized rather than squared daily returns; Andersen et al. (2001b,a) documented the distributional properties of realized variance for exchange rates and equities; and Barndorff-Nielsen and Shephard (2002) developed its asymptotic theory. Andersen et al. (2003) demonstrated that simple reduced-form models of realized volatility outperform latent-variable GARCH and stochastic volatility models, the observation that motivates the entire measure-then-model paradigm in which this paper operates. Andersen et al. (2005) further showed that evaluating forecasts against a realized proxy, rather than against squared returns, removes most of the apparent "unforecastability" of volatility.

At very high sampling frequencies, market microstructure noise breaks the consistency of plain realized variance (Hansen and Lunde, 2006b; Bandi and Russell, 2008; Aït-Sahalia et al., 2005). The literature responded with noise-robust estimators: subsampled and two-scale estimators (Zhang et al., 2005) and realized kernels (Barndorff-Nielsen et al., 2008, 2009). Liu et al. (2015) compared some 400 realized measures across 31 instruments from several asset classes and found that 5-minute RV is remarkably hard to beat, a finding that guides our choice of subsampled 5-minute RV as the headline measure, with 1-minute RV and realized kernels as robustness checks. A parallel branch separates the continuous and jump components of quadratic variation: bipower variation (Barndorff-Nielsen and Shephard, 2004), formal jump tests (Barndorff-Nielsen and Shephard, 2006), and threshold refinements (Corsi et al., 2010). Barndorff-Nielsen et al. (2010) introduced realized semivariances, whose signed decomposition Patton and Sheppard (2015) showed to matter for persistence: negative semivariance predicts future volatility far more strongly than positive semivariance. The quality of the underlying tick data is a first-order concern for all of these measures (Brownlees and Gallo, 2006), and Alexander and Dakos (2020) document how severe the problem is for cryptocurrency data, where exchanges differ in liquidity, data integrity and trading hours.

2.2. HAR models and extensions

Corsi (2009) proposed the heterogeneous autoregressive (HAR) model, whose daily–weekly–monthly cascade approximates long memory with three OLS coefficients, following the heterogeneous-market intuition of Müller et al. (1997); it has become the workhorse benchmark of the field (Clements and Preve, 2021). Extensions incorporate the continuous/jump split (Andersen et al., 2007), signed semivariances (Patton and Sheppard, 2015), and attenuation-bias corrections: Bollerslev et al. (2016) let the daily coefficient shrink on days when realized quarticity signals a noisy RV estimate (HARQ), and Buccheri and Corsi (2021) treat the measurement error explicitly in a state-space form. Logarithmic specifications are common because log RV is close to Gaussian (Andersen et al., 2001b) and the log transform tames heteroskedastic measurement error. On the GARCH side, Hansen and Lunde (2005) showed that little beats a GARCH(1,1) on daily exchange-rate returns once forecasts are evaluated against a realized proxy, while Hansen et al. (2012) and Hansen and Huang (2016) tied the latent-variance recursion to a realized measure (Realized GARCH and Realized EGARCH), typically closing most of the gap to HAR-type models. Our benchmark suite spans exactly this spectrum—six HAR specifications, three daily GARCH variants, and the Realized GARCH—so that the machine-learning comparison is anchored against the strongest available econometrics rather than against a single straw man.

2.3. Machine learning for volatility forecasting

Applications of statistical learning to realized volatility are more recent, and the evidence is mixed. Audrino and Knaus (2016) used the LASSO to select lag structures and found that it recovers HAR-like patterns. Bucci (2020) documented gains from recurrent neural networks for monthly equity-index volatility. Rahimikia and Poon (2020) trained learners on large limit-order-book and news datasets for intraday volatility. Closest to our design, Christensen et al. (2023) ran a systematic horse race of regularized regressions, trees and neural networks against HAR on 29 Dow Jones stocks and found double-digit QLIKE improvements concentrated in periods of market stress and at longer horizons; Zhang et al. (2024) showed that much of the gain of nonlinear learners on US equities comes from pooling across stocks and exploiting intraday commonality. Against these positive findings, Branco et al. (2024) report that, on a broad panel of equities, carefully specified linear HAR-type models are not beaten by random forests, boosting or neural networks once the comparison is conducted with robust losses and proper inference. In the broader return-prediction literature, Gu et al. (2020) established the now-standard result that trees and shallow networks dominate linear methods and that pooling across assets is crucial for regularization—a theme that our pooled multi-asset model imports into the volatility context. For cryptocurrencies, GARCH-type comparisons dominate (Katsiampa, 2017), model and parameter instability is pervasive (Catania et

al., 2019), and high-frequency machine-learning applications remain scarce.

2.4. Forecast evaluation and economic value

Because volatility is observed only through noisy proxies, loss functions must be chosen carefully: Hansen and Lunde (2006a) and Patton (2011) characterize the class of robust losses, of which QLIKE and MSE are the leading members, for which rankings under a proxy coincide with rankings under the true conditional variance, and Patton (2020) shows that the same consistency carries over to the comparison of possibly misspecified forecasts. Statistical comparisons rely on the Diebold and Mariano (1995) test, whose use for comparing forecasting *methods* estimated on rolling windows is justified by Giacomini and White (2006) (see also West, 1996; Diebold, 2015), on the model confidence set of Hansen et al. (2011), and, in unstable environments, on the fluctuation analysis of Giacomini and Rossi (2010). Forecast combination is a persistent empirical winner (Bates and Granger, 1969; Timmermann, 2006). The economic-value strand converts statistical accuracy into portfolio outcomes: Fleming et al. (2001, 2003) priced volatility timing with mean-variance utility, Bollerslev et al. (2018) showed on more than fifty futures markets that realized-volatility-based risk models deliver substantial utility gains in a risk-targeting framework, and risk-management evaluation backtests VaR coverage (Kupiec, 1995; Christoffersen, 1998). Patton and Sheppard (2009) survey the evaluation toolkit.

2.5. Contribution relative to the literature

Three gaps motivate this paper. First, the ML-versus-HAR evidence comes almost exclusively from single asset classes—typically US equities (Christensen et al., 2023; Zhang et al., 2024; Branco et al., 2024)—and it is contradictory even there, which leaves open whether any documented gain generalizes to FX, futures and crypto, whose volatility dynamics, trading sessions and noise properties differ sharply. Second, the pooling and *transferability* question—whether one model trained on many assets can forecast an asset it has never seen—has been studied for returns (Gu et al., 2020) and for equity volatility (Zhang et al., 2024), but not across asset classes. Third, statistical gains are rarely pushed through to economic value and risk management within the same controlled design. We address all three within a single, fully reproducible pipeline: identical rolling protocol, identical robust losses, identical inference, from raw 1-minute bars to utility gains and VaR coverage, across 25 assets in four classes.

3. Data

3.1. Source and universe

All price data come from the HF Market Data API (https://www.hfmarketdata.io), an open high-frequency data service providing intraday and daily OHLCV bars for equities, ETFs, continuous futures, cryptocurrencies, indices and FX. We download 1-minute and daily bars from January 2013 through July 2026 for a universe of 25 forecast targets spanning four asset classes, plus the VIX as a predictor. The *equity* class contains five ETFs (SPY, QQQ, IWM, GLD, TLT) and five large-capitalization stocks (AAPL, MSFT, NVDA, JPM, XOM), all split- and dividend-adjusted. The *foreign exchange* class contains five major pairs (EURUSD, USDJPY, GBPUSD, USDCAD, AUDUSD). The *cryptocurrency* class contains BTC, ETH, SOL, ADA and XRP.[1] The *futures* class contains five continuous front-month series (ES, S&P 500; NQ, Nasdaq 100; CL, WTI crude; GC, gold; ZN, 10-year Treasury note), ratio back-adjusted so that roll dates do not generate spurious returns. The universe is deliberately tilted towards the most liquid instrument of each class, so that the comparison is not driven by illiquidity, while remaining large enough for cross-sectional inference on model rankings.

In total the pipeline processes 93.8 million 1-minute bars and produces a panel of 84,898 asset-day observations of realized measures. A reusable API client handles rate limiting, exponential back-off retries, pagination and a local parquet cache, making the entire pipeline reproducible end to end from raw bars.

3.2. Trading sessions and data handling

Timestamps are served in exchange-local (US Eastern) time for every instrument. For equities and ETFs, whose 1-minute bars cover the extended session (04:00–20:00), we retain regular trading hours only (09:30–16:00) when computing realized measures, so that RV is not contaminated by thin pre- and post-market trading. FX trades essentially continuously from Sunday 17:00 to Friday 17:00; its trading day is the calendar day, and weekend fragments are eliminated by the minimum-bar filter. Futures trade nearly 24 hours with short maintenance breaks; their day is likewise the calendar day. Crypto trades continuously; its day is the calendar day in US Eastern time, as served by the provider, so that a crypto "day" closes four hours after the US equity session and weekends are genuine trading days—crypto therefore contributes roughly 7/5 as many observations per calendar year as the session-limited markets. Days with fewer than 700 valid 1-minute bars (250 for the 6.5-hour equity session)—holidays, half-days and outage days—are discarded. Missing

[1] The originally targeted BNB is not available in the data source and is replaced by ADA. ETH, XRP, ADA and SOL enter the sample at their first available dates (May 2016, March 2017, January 2019 and April 2021, respectively); BTC is available from April 2013.

minutes within a day (no trade) are handled by computing returns between consecutive observed bars, and the 5-minute grids are built on observed prices, so no artificial zero returns are introduced. No further cleaning of the bars is applied; we return to the consequences of this choice for the early crypto sample in Section 8.

3.3. Descriptive statistics

Table 1 reports per-instrument descriptive statistics of the daily realized measures; Figure 1 plots the annualized RV series for one representative asset per class, and Figure 3 the autocorrelation functions of log RV. Three features of the data matter for the forecasting exercise. First, the levels differ enormously across classes: mean annualized volatility is 7.9% for FX, 15.9% for equities and ETFs, 16.8% for futures, and 108% for crypto (median 68.9%)—any pooled model must absorb these level differences before it can share dynamics. Second, persistence is uniformly high—the first autocorrelation of log RV averages 0.70 for FX, 0.74 for equities, 0.77 for futures and 0.86 for crypto—and decays slowly (Figure 3), the signature of long memory that motivates the HAR cascade. Third, jump days identified by the BNS test at $\alpha = 0.1\%$ range from 3.7% of days for equities to 8.2% for FX, 10.2% for futures and 14.8% for crypto, so jump-robust measures and jump-aware models are potentially material outside equities.

3.4. Microstructure noise across measures

Figure 2 compares yearly averages of 1-minute RV, subsampled 5-minute RV and the realized kernel for SPY and BTC. For SPY the three measures are nearly indistinguishable—at SPY's liquidity, 1-minute sampling carries little noise, and the pairwise correlations between measures exceed 0.98. For BTC the ordering $RV^{(1\text{min})} > RV^{ss} > RK$ in the early sample (2013–2015) is exactly the pattern implied by i.i.d. microstructure noise, whose variance contribution grows linearly in the sampling frequency (Hansen and Lunde, 2006b): early Bitcoin trading was thin and fragmented, and plain 1-minute RV overstates variance by a factor of two or more relative to the noise-robust kernel. The gap closes progressively as crypto market quality improves after 2017. This pattern justifies our choice of the subsampled 5-minute RV as the headline measure—following the multi-asset evidence of Liu et al. (2015)—with the 1-minute RV and the realized kernel used to assess robustness to the sampling scheme in Section 6. It also means that the crypto series contain a handful of days on which even the subsampled 5-minute RV is implausibly large (annualized volatility above 1,000%, concentrated in 2013 for BTC and in 2017 and early 2021 for the altcoins); these days are retained, because the evaluation losses we use are robust to noise in the proxy (Patton, 2011), but their effect on the crypto results is examined in Section 6.

4. Methodology

4.1. Realized measures

Let $p_{t,i}$ denote the i-th intraday log price of a given asset on trading day t, sampled on an equidistant grid of n intraday returns $r_{t,i} = p_{t,i} - p_{t,i-1}$. Under the standard jump–diffusion semimartingale

$$dp_s = \mu_s\, ds + \sigma_s\, dW_s + \kappa_s\, dq_s, \tag{1}$$

where W is a Brownian motion, q a counting process, and κ the jump size, the realized variance

$$RV_t \;=\; \sum_{i=1}^{n} r_{t,i}^2 \;\xrightarrow{\;p\;}\; \underbrace{\int_{t-1}^{t} \sigma_s^2\, ds}_{\text{integrated variance}} \;+\; \underbrace{\sum_{t-1<s\le t} \kappa_s^2}_{\text{jump variation}} \tag{2}$$

consistently estimates total quadratic variation as $n \to \infty$ (Andersen and Bollerslev, 1998; Barndorff-Nielsen and Shephard, 2002). In practice market microstructure noise contaminates returns at the very highest frequencies (Hansen and Lunde, 2006b; Bandi and Russell, 2008), and we therefore compute several noise-robust variants.

Subsampled 5-minute RV. Our headline measure, denoted RV_t^{ss}, averages the realized variance computed on five staggered 5-minute grids offset by one minute each,

$$RV_t^{ss} \;=\; \frac{1}{5} \sum_{k=0}^{4} \sum_{j} \left(p_{t,5j+k} - p_{t,5(j-1)+k} \right)^2, \tag{3}$$

in the spirit of the subsampling scheme of Zhang et al. (2005). Following Liu et al. (2015), 5-minute sampling is a demanding benchmark that is difficult to beat across asset classes; subsampling recovers part of the efficiency lost to sparse sampling.

Realized kernel. We also compute the non-flat-top Parzen realized kernel of Barndorff-Nielsen et al. (2008),

$$RK_t \;=\; \gamma_0 + \sum_{h=1}^{H} k\!\left(\tfrac{h}{H+1}\right) (\gamma_h + \gamma_{-h}), \qquad \gamma_h = \sum_{i} r_{t,i}\, r_{t,i-h}, \tag{4}$$

on 1-minute returns, where $k(\cdot)$ is the Parzen kernel and the bandwidth follows the feasible rule $H^* = c^* \xi^{4/5} n^{3/5}$ with $c^* = 3.5134$, noise-to-signal ratio $\xi^2 = \hat{\omega}^2 / \widehat{IV}$, noise variance estimated as $\hat{\omega}^2 = RV_t^{(1\text{min})}/(2n)$, and $\widehat{IV}$ proxied by RV_t^{ss} (Barndorff-Nielsen et al., 2009). The non-flat-top Parzen weights guarantee non-negativity.

Bipower variation and jumps. The realized bipower variation of Barndorff-Nielsen and Shephard (2004),

$$BV_t \;=\; \mu_1^{-2}\,\frac{n}{n-1}\sum_{i=2}^{n}|r_{t,i}|\,|r_{t,i-1}|, \qquad \mu_1=\sqrt{2/\pi}, \tag{5}$$

is consistent for integrated variance in the presence of jumps. Jump days are identified with the ratio-form test of Barndorff-Nielsen and Shephard (2006),

$$Z_t \;=\; \frac{1-BV_t/RV_t}{\sqrt{\theta\,\max\!\left(1,\,TQ_t/BV_t^2\right)/n}} \;\xrightarrow{d}\; N(0,1), \qquad \theta=\tfrac{\pi^2}{4}+\pi-5, \tag{6}$$

where the tripower quarticity

$$TQ_t \;=\; n\,\mu_{4/3}^{-3}\,\frac{n}{n-2}\sum_{i=3}^{n}|r_{t,i}|^{4/3}\,|r_{t,i-1}|^{4/3}\,|r_{t,i-2}|^{4/3}, \qquad \mu_{4/3}=2^{2/3}\,\Gamma(7/6)/\Gamma(1/2), \tag{7}$$

is a jump-robust estimator of integrated quarticity. The jump and continuous components are then

$$J_t=\mathbb{1}\{Z_t>\Phi_{1-\alpha}^{-1}\}\cdot\max(RV_t-BV_t,0), \qquad C_t=RV_t-J_t, \tag{8}$$

with $\alpha=0.001$ (Andersen et al., 2007). To limit noise distortions, BV_t, TQ_t and the test are computed on the sparse 5-minute grid (offset zero), so that RV_t in (6)–(8) is the plain 5-minute RV of that grid.

Realized semivariances and quarticity. Following Barndorff-Nielsen et al. (2010), the positive and negative semivariances

$$RS_t^{+}=\sum_i r_{t,i}^2\,\mathbb{1}\{r_{t,i}>0\}, \qquad RS_t^{-}=\sum_i r_{t,i}^2\,\mathbb{1}\{r_{t,i}<0\}, \tag{9}$$

decompose the 5-minute RV_t by the sign of the underlying returns; the signed jump variation is $SJ_t=RS_t^{+}-RS_t^{-}$ (Patton and Sheppard, 2015). Finally, the realized quarticity $RQ_t=\frac{n}{3}\sum_i r_{t,i}^4$, also on the 5-minute grid, estimates the sampling variability of RV_t and feeds the HARQ model below.

Trading sessions. Equity and ETF measures use regular trading hours only (09:30–16:00, exchange time); FX and futures use the full trading day on a calendar-day partition, with holidays and thin days (fewer than 700 one-minute bars, 250 for equities) discarded; crypto trades continuously and its day is the calendar day in US Eastern time. Daily close-to-close log returns r_t, which include the overnight component for session-limited markets, feed the GARCH-family models and the economic-value exercises.

4.2. Forecast target and rolling protocol

For horizons $h \in \{1, 5, 22\}$ trading days the target is the *average* daily realized variance over the next h days,

$$\overline{RV}_{t+1:t+h} \;=\; \frac{1}{h}\sum_{j=1}^{h} RV^{ss}_{t+j}, \tag{10}$$

so that cumulative variance equals h times the target. Every model produces a forecast $F_{t,h}$ of (10) using information available at the close of day t. All models are estimated on a rolling window of 1,000 trading days of the asset's own calendar and re-estimated every 22 trading days; forecasts are produced daily. For direct multi-step projections the estimation sample at origin t contains only origins $s \le t - h$, whose targets are fully realized by t, so that no future information enters any parameter estimate. The evaluation sample of each asset therefore starts after its first $1{,}000 + h$ observations—December 2016 for the series that begin in 2013, later for the altcoins—and runs through July 2026. For each asset and horizon, all models are evaluated on the *common* set of dates on which every model produced a forecast, so that average losses, pairwise tests and the model confidence set are computed on identical samples.

4.3. Econometric benchmarks

HAR family. The heterogeneous autoregressive model of Corsi (2009) regresses the target on daily, weekly and monthly averages,

$$\overline{RV}_{t+1:t+h} \;=\; \beta_0 + \beta_d\, RV_t + \beta_w\, \overline{RV}_{t-4:t} + \beta_m\, \overline{RV}_{t-21:t} + \varepsilon_{t+h}, \tag{11}$$

estimated by OLS separately for each horizon (direct projection). We consider five extensions, six HAR specifications in total: HAR-J adds the jump component J_t; HAR-CJ replaces the RV averages with continuous-component averages C_t, $\overline{C}_{t-4:t}$, $\overline{C}_{t-21:t}$ and adds J_t (Andersen et al., 2007); SHAR splits the daily lag into RS^{+}_t and RS^{-}_t (Patton and Sheppard, 2015); HARQ adds the interaction $RV_t\, RQ_t^{1/2}$ so that the effective daily coefficient $\beta_d + \beta_Q\, RQ_t^{1/2}$ shrinks towards the longer-run averages on days when RV_t is measured imprecisely (Bollerslev et al., 2016); and log-HAR estimates (11) in logarithms, with forecasts retransformed as $\exp(\hat{y} + \hat{\sigma}^2_\varepsilon/2)$ under lognormality of the residual. Following Bollerslev et al. (2016), all forecasts of all models pass through an "insanity filter" that clips them to the range of the target observed in the estimation window.

GARCH family. On daily returns $r_t = \mu + \varepsilon_t$, $\varepsilon_t = \sigma_t z_t$, $z_t \sim N(0,1)$, we estimate by Gaussian quasi-maximum likelihood the GARCH(1,1) of Bollerslev (1986),

$$\sigma^2_t = \omega + \alpha\, \varepsilon^2_{t-1} + \beta\, \sigma^2_{t-1}, \tag{12}$$

the GJR-GARCH of Glosten et al. (1993), which adds the leverage term $\gamma \varepsilon_{t-1}^2 \mathbb{1}\{\varepsilon_{t-1} < 0\}$ to (12), and the EGARCH of Nelson (1991),

$$\ln \sigma_t^2 = \omega + \alpha\big(|z_{t-1}| - \mathbb{E}|z_{t-1}|\big) + \gamma z_{t-1} + \beta \ln \sigma_{t-1}^2. \tag{13}$$

Because these models forecast the conditional variance of close-to-close returns, which includes overnight variation absent from intraday RV, their k-step variance forecasts $\hat{\sigma}_{t+k|t}^2$ are mapped into RV units with the proportionality factor

$$\hat{c}_t = \frac{W^{-1}\sum_{s=t-W+1}^{t} RV_s^{ss}}{W^{-1}\sum_{s=t-W+1}^{t} \hat{\varepsilon}_s^2}, \qquad F_{t,h}^{\mathrm{GARCH}} = \hat{c}_t \, \frac{1}{h}\sum_{k=1}^{h} \hat{\sigma}_{t+k|t}^2, \tag{14}$$

re-estimated on each window of $W = 1{,}000$ days, in the spirit of the realized-proxy evaluation of Hansen and Lunde (2005). Parameters are re-estimated every 22 days, but the variance recursion is re-filtered every day on the data through t with the current parameters, so that (14) conditions on the same information as the HAR forecasts.[2] Multi-step variance paths are analytic for GARCH and GJR and simulated (500 paths, fixed seed) for EGARCH.

Realized GARCH. The log-linear Realized GARCH(1,1) of Hansen et al. (2012) couples the latent variance h_t with the realized measure $x_t = RV_t^{ss}$:

$$\begin{aligned} r_t &= \sqrt{h_t}\, z_t, \qquad z_t \sim N(0,1), \\ \log h_t &= \omega + \beta \log h_{t-1} + \gamma \log x_{t-1}, \\ \log x_t &= \xi + \varphi \log h_t + \tau_1 z_t + \tau_2 (z_t^2 - 1) + u_t, \qquad u_t \sim N(0, \sigma_u^2). \end{aligned} \tag{15}$$

We estimate (15) by quasi-maximum likelihood and forecast the realized measure directly. Substituting the measurement equation into the GARCH equation gives the reduced-form recursion $\log h_{t+1} = (\omega + \gamma\xi) + \pi \log h_t + \gamma \eta_t$, with persistence $\pi = \beta + \gamma\varphi$ and innovation $\eta_t = \tau_1 z_t + \tau_2(z_t^2 - 1) + u_t$ of variance $\sigma_\eta^2 = \tau_1^2 + 2\tau_2^2 + \sigma_u^2$; the k-step forecast is $\mathbb{E}_t[x_{t+k}] = \exp\big(\xi + \varphi\, \mathbb{E}_t[\log h_{t+k}] + \tfrac{1}{2}(\varphi^2 V_{t,k} + \sigma_\eta^2)\big)$, where $V_{t,k}$ is the accumulated forecast variance of $\log h_{t+k}$, so that the lognormal correction is explicit.

4.4. Machine-learning models

Information sets. All learners forecast $\log \overline{RV}_{t+1:t+h}$ and are evaluated after retransformation with the nonparametric smearing factor of Duan (1983), $\hat{F} = \exp(\hat{y}) \cdot N^{-1}\sum_s \exp(\hat{e}_s)$ with $\hat{e}_s$ the

[2] The first version of our code produced the GARCH forecast only at re-estimation dates and carried it forward unchanged for the next 22 days; the forecasts reported here correct this.

in-window residuals, and the same insanity filter as the HAR models. Two nested predictor sets are used. The *HAR set* contains exactly the HAR information: $\log RV_t$, $\log \overline{RV}_{t-4:t}$, $\log \overline{RV}_{t-21:t}$. The *extended set* adds fifteen predictors, all known at the end of day t: the jump share J_t/RV_t, the log semivariances, the normalized signed jump SJ_t/RV_t, the log realized quarticity, the daily, weekly and monthly close-to-close returns, the log volume growth, the day of the week, the log VIX level, and the cross-sectional averages of log RV within each of the four asset classes. The cross-asset averages are contemporaneous: since every class closes at or before the crypto day's end, they are available to a crypto forecaster at the close of its day, while for the session-limited markets the crypto average refers to the (partly overlapping) calendar day.

Learners. We consider penalized linear models—ridge, LASSO (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) on standardized predictors—and tree ensembles: random forests (Breiman, 2001), XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017). Hyperparameters are tuned by walk-forward validation inside the estimation window (the last 20% of the window, in time order, serves as the validation split for a small grid), re-tuned once a year, and the winning configuration is refit on the full window; no future information enters the tuning loop at any point.

Deep learners. Two sequence models map the last 22 days of extended features to all three horizons jointly: a single-layer LSTM (Hochreiter and Schmidhuber, 1997) with 64 hidden units, and a two-block Transformer encoder (Vaswani et al., 2017) with model dimension 32 and 4 attention heads. Both are trained with Adam, early stopping on a temporal validation split (the last 15% of the window), and quarterly re-estimation (every 66 trading days) to keep the computational burden manageable.

Pooled multi-asset model. Beyond the asset-by-asset models, a single pooled LightGBM is trained on the stacked panel of all 25 assets with the ticker and asset-class identifiers as categorical features, re-estimated every 22 days of the union calendar on the trailing 1,000-day window. The pooled design tests whether volatility dynamics are sufficiently common across assets for cross-sectional information to sharpen asset-level forecasts (Gu et al., 2020; Christensen et al., 2023; Zhang et al., 2024). A leave-one-out variant excludes a target asset from the training pool entirely and predicts it out-of-pool, measuring the transferability of pooled dynamics to unseen assets. Model interpretation relies on SHAP values (Lundberg and Lee, 2017) and permutation importance.

4.5. Forecast evaluation

Loss functions. Our primary loss is the QLIKE,

$$L^{Q}(RV, F) \;=\; \frac{RV}{F} - \log\frac{RV}{F} - 1, \tag{16}$$

which together with the MSE, $L^{M}(RV,F) = (RV - F)^2$, belongs to the class of losses that are robust to noise in the volatility proxy (Hansen and Lunde, 2006a; Patton, 2011): rankings based on (16) with RV in place of the latent variance are consistent for the rankings that would obtain under the true conditional variance. QLIKE is also scale-free, which permits averaging across assets whose variance levels differ by two orders of magnitude; MSE results are reported as a complement. We report mean losses and, for every model m, the loss ratio $\bar{L}^{Q}_{m}/\bar{L}^{Q}_{\mathrm{HAR}}$ relative to the HAR benchmark.

Pairwise tests. For two models with per-period losses $L_{a,t}$ and $L_{b,t}$ and $d_t = L_{a,t} - L_{b,t}$, the Diebold and Mariano (1995) statistic is

$$DM \;=\; \frac{\bar{d}}{\sqrt{\hat{\Omega}/T}}, \qquad \hat{\Omega} = \hat{\gamma}_0 + 2\sum_{k=1}^{h-1}\left(1 - \tfrac{k}{h}\right)\hat{\gamma}_k, \tag{17}$$

where $\hat{\gamma}_k$ is the k-th sample autocovariance of d_t and the Bartlett bandwidth $h-1$ (Newey and West, 1987) accounts for the overlap of h-step forecast errors. Since all models are estimated on rolling windows of fixed length, the statistic compares forecasting *methods* rather than population models, and the asymptotic normality of (17) follows from Giacomini and White (2006) rather than from West (1996). Negative values favour model a. We run (17) for every model against HAR, asset by asset, and summarize the 25 statistics per model and horizon by the share of assets on which the model is significantly better or worse than HAR at the 5% level.

Model confidence set. Joint comparisons use the model confidence set of Hansen et al. (2011). For the set $\mathcal{M}$ of surviving models let $\bar{d}_{i\cdot} = \bar{L}_i - |\mathcal{M}|^{-1}\sum_{j\in\mathcal{M}}\bar{L}_j$ be the deviation of model i's mean loss from the set average, and

$$t_{i\cdot} = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\mathrm{var}}(\bar{d}_{i\cdot})}}, \qquad T_{\max} = \max_{i\in\mathcal{M}} t_{i\cdot}, \tag{18}$$

where the variance is estimated from $B = 5{,}000$ stationary-bootstrap replications (Politis and Romano, 1994) of the loss series with mean block length 10. The distribution of $T_{\max}$ under the null of equal predictive ability is obtained from the same replications; if the null is rejected at level α the model with the largest $t_{i\cdot}$ is eliminated and the procedure iterates. The MCS p-value of a

model is the largest test p-value up to and including its elimination, and the 90% MCS collects the models with p-value at least 0.10. The MCS is computed asset by asset and horizon by horizon on the common sample, and we report the share of assets for which each model belongs to the set.

Calibration and encompassing. Forecast unbiasedness is assessed with Mincer and Zarnowitz (1969) levels regressions,

$$\overline{RV}_{t+1:t+h} = a + b\,F_{t,h} + e_{t+h}, \tag{19}$$

testing $(a,b) = (0,1)$ with a Wald statistic based on HAC standard errors (bandwidth $h-1$). Whether a rival forecast contains information absent from HAR is tested with the forecast-encompassing regression of Harvey et al. (1998),

$$\overline{RV}_{t+1:t+h} = a + b_1 F^{\mathrm{HAR}}_{t,h} + b_2 F^{m}_{t,h} + e_{t+h}; \tag{20}$$

rejection of $b_2 = 0$ (HAC t-test) with $b_2 > 0$ means that HAR does not encompass model m.

Forecast combinations. We report two combinations (Bates and Granger, 1969; Timmermann, 2006): the simple mean of all individual forecasts, and inverse-MSE weights

$$w_{m,t} = \frac{\widehat{\mathrm{MSE}}_{m,t}^{-1}}{\sum_j \widehat{\mathrm{MSE}}_{j,t}^{-1}}, \qquad \widehat{\mathrm{MSE}}_{m,t} = \frac{1}{60}\sum_{s=t-60}^{t-1}\left(\overline{RV}_{s+1:s+h} - F^{m}_{s,h}\right)^2, \tag{21}$$

computed on a trailing 60-day window of realized squared errors, so that the weights use only information available at t.

Stability. Because relative performance may drift across regimes, we complement full-sample tests with the fluctuation test of Giacomini and Rossi (2010), which tracks the standardized rolling-window DM statistic

$$\mathscr{F}_t = \frac{\sqrt{m}\,\bar{d}_t^{(m)}}{\hat{\sigma}_d}, \qquad \bar{d}_t^{(m)} = \frac{1}{m}\sum_{s=t-m+1}^{t} d_s, \tag{22}$$

over a window of $m = 0.3\,T$ observations, with $\hat{\sigma}_d$ the full-sample HAC standard deviation of d_t, and compares its path with the regime-robust critical values tabulated by Giacomini and Rossi (2010) for $\mu = m/T$.

4.6. Economic value

Volatility timing. For each asset and model we run the volatility-timing strategy of Fleming et al. (2001, 2003): the weight on the risky asset is

$$w_t = \min\left(\bar{w},\ \frac{\sigma^*}{\sqrt{252\,F_{t,1}}}\right), \tag{23}$$

with annualized target $\sigma^* = 10\%$, leverage cap $\bar{w} = 3$, and $F_{t,1}$ the one-day-ahead variance forecast; the weight chosen at the close of t is applied to the close-to-close return of $t+1$, and net returns subtract proportional transaction costs of 5 basis points on turnover $|w_t - w_{t-1}|$ (0–20 bps in sensitivity checks). We report annualized Sharpe ratios and the Fleming et al. (2001) utility gain: for a mean-variance investor with relative risk aversion $\gamma = 5$ and utility $U(r) = \mathbb{E}[r] - \frac{\gamma}{2}\mathrm{Var}[r]$ on daily net returns, the annualized fee

$$\Delta_m \;=\; 252\left[U(r^m) - U(r^{\mathrm{HAR}})\right] \tag{24}$$

is the amount the investor would pay to switch from HAR-based to model-m-based timing. Because the pooled strategies are driven by variance forecasts of the *intraday* RV, which omits overnight variance, the realized strategy volatility differs from the target; the comparison across models is nevertheless like for like.

Value-at-Risk. One-day Gaussian VaR forecasts, $\mathrm{VaR}^q_{t+1} = -z_q\,\hat{\sigma}^r_{t+1}$, are backtested at the 1% and 5% levels. Because RV omits overnight variance, forecasts are rescaled into close-to-close return units with the trailing 250-day ratio of mean squared returns to mean forecast variance, so no future information is used. With x violations in T days and $\hat{\pi} = x/T$, the unconditional-coverage test of Kupiec (1995) is

$$LR_{uc} = -2\left[(T-x)\ln(1-q) + x\ln q\right] + 2\left[(T-x)\ln(1-\hat{\pi}) + x\ln\hat{\pi}\right] \;\sim\; \chi^2_1, \tag{25}$$

and the conditional-coverage test of Christoffersen (1998) adds the likelihood-ratio statistic for first-order independence of the violation indicator, $LR_{cc} = LR_{uc} + LR_{ind} \sim \chi^2_2$. We report mean violation rates and the share of assets for which each model passes both tests at the 5% level.

5. Results

5.1. Evaluation sample

The common evaluation sample contains 59,323 asset-days at the one-day horizon (59,201 at five days and 58,776 at 22 days), each with 27 forecasts: the six HAR specifications, the four GARCH-type models, twelve tabular learners, the two deep learners, the pooled LightGBM and the two combinations. For the FX pairs and the futures the sample starts in December 2016, for the equities and ETFs in January or February 2017, for ETH in September 2019, for XRP in February 2020, for ADA in November 2021 and for SOL in February 2024 (894 days), and it ends on 29 July 2026 for $h = 1$ (Table 12 in the appendix lists the per-asset counts). All statistics below are computed on this common sample. Mean losses are first computed asset by asset and then averaged across the 25 assets, so that the crypto series, which have more trading days, do not receive a larger weight than the others; the sub-period and sampling-frequency tables of Section 6 pool asset-days instead and are flagged as such.

5.2. Average losses

Table 2 reports the mean QLIKE of each model and its ratio to the HAR benchmark at the three horizons, and Figure 4 plots the ratios. Four results stand out.

First, the HAR model in levels is a weak benchmark on this cross-section: its mean QLIKE is 0.256, 0.271 and 0.288 at $h = 1$, 5 and 22, and almost every model that forecasts the *logarithm* of realized variance beats it by a wide margin. Log-HAR, which uses exactly the same three predictors, attains ratios of 0.75, 0.63 and 0.79. The ridge, LASSO and elastic-net regressions on the HAR information set, which differ from log-HAR only through the penalty, obtain almost identical ratios (0.75, 0.63 and 0.80 for ridge). The bulk of the "machine-learning gain" documented below is therefore the gain from modelling log variance with a lognormal retransformation, not from nonlinearity or from richer information.

Second, within the log-linear family the extended information set adds a further, smaller improvement: ridge on the extended set reaches 0.70, 0.61 and 0.79, elastic net 0.70, 0.61 and 0.79, LASSO 0.71, 0.61 and 0.80. The extended set thus buys about five QLIKE points at the short horizons and nothing at one month. The penalized linear models on the extended set are the best individual models at $h = 1$ and $h = 5$; the inverse-MSE combination is marginally better at $h = 22$ (0.77) and within one point of the best at the shorter horizons (0.72 and 0.61). Asset by asset, ridge on the extended set has the lowest one-day QLIKE for 13 of the 25 instruments and elastic net for 7; a tree ensemble is the best model for only two instruments and the pooled model for one.

Third, the tree ensembles and the deep learners do not keep up with the linear learners. Random

forests on the extended set reach 0.79, 0.72 and 0.95; XGBoost 0.80, 0.80 and 1.12; LightGBM 0.85, 0.85 and 1.16. At one month every boosted model is worse than the HAR benchmark in levels. The LSTM (1.33, 1.05, 1.05) and the Transformer (1.89, 1.48, 1.46) are worse than HAR at every horizon. Pooling across assets helps the trees considerably: the pooled LightGBM reaches 0.77, 0.70 and 0.95, beats the asset-by-asset LightGBM on 23 of 25 instruments at $h = 1$, and is the only tree-based model to rival the log-linear regressions at the short horizons; it does not, however, beat log-HAR on more than 9 of the 25 assets.

Fourth, on the econometric side the Realized GARCH is the only GARCH-type model that improves on HAR (0.78, 0.66 and 0.88), and it beats the daily GARCH(1,1) on all 25 assets, in line with Hansen et al. (2012). The daily GARCH, GJR and EGARCH models, which see only close-to-close returns, are 12 to 57 percent worse than HAR. Among the HAR extensions, the jump models HAR-J and HAR-CJ gain about six points at $h = 1$ and twelve at $h = 5$ but lose at $h = 22$, SHAR is indistinguishable from HAR, and HARQ is the worst model in the table. The HARQ result deserves a comment. On SPY, for instance, the median QLIKE of HARQ (0.112) is lower than that of HAR (0.133), as in Bollerslev et al. (2016), but its mean (0.432 against 0.249) is dominated by a small number of days on which the realized-quarticity interaction drives the forecast far below the realized value—five percent of its SPY forecasts are less than half the HAR forecast—and the resulting QLIKE losses are enormous (a maximum of 272). The insanity filter of Bollerslev et al. (2016), which clips forecasts to the range of the target in the estimation window, is not restrictive enough to contain this behaviour when the realized quarticity is estimated from 78 to 288 five-minute returns, and the problem is most acute for the futures (ratio 3.29) and the cryptocurrencies (1.92).

5.3. Results by asset class

Table 3 breaks the one-day and one-month ratios down by asset class. The ordering of the models is remarkably stable across classes, but the size of the gains is not. At $h = 1$ the best linear learner (ridge, extended set) attains 0.62 for crypto, 0.72 for equities and ETFs, 0.78 for futures and 0.85 for FX; log-HAR follows the same gradient (0.65, 0.76, 0.88, 0.91). The gains over HAR are thus largest where volatility is most volatile and most skewed—the cryptocurrencies—and smallest in FX, where realized variance is the smoothest and the HAR benchmark in levels is already close to the best available forecast. The tree ensembles reverse part of the picture: LightGBM on the extended set is competitive only for crypto (0.67), is roughly on par with HAR for equities (0.95) and is worse than HAR for futures and FX (1.03 and 1.02); at one month it is worse than HAR in every class (0.94 to 1.38). The GARCH models are worse than HAR in every class at $h = 1$, most dramatically for crypto (1.60), but at one month they beat HAR for equities and futures (0.92 and 0.91), where the overnight component they capture matters more for the monthly variance than the

intraday precision they lack. The Realized GARCH is between 0.71 and 0.94 in every class at both horizons, which confirms that the realized measure, rather than the GARCH structure, is what the daily models were missing.

5.4. Pairwise tests and model confidence sets

Table 4 summarizes the asset-by-asset Diebold–Mariano tests against HAR and the 90% model confidence sets, and Figure 5 plots the MCS inclusion shares. At $h = 1$ the penalized linear models on the extended set and the inverse-MSE combination are significantly better than HAR at the 5% level on 24 of the 25 instruments and significantly worse on none; log-HAR and ridge on the HAR set are significantly better on 21 and 22 instruments; the Realized GARCH on 17; the pooled LightGBM on 15. For ridge on the extended set the result holds in every class (10 of 10 equities, 5 of 5 FX pairs, 5 of 5 futures, 4 of 5 cryptocurrencies). The asset-by-asset LightGBM, by contrast, is significantly better than HAR on 4 instruments and significantly worse on 8, the GARCH(1,1) is worse on 20 and the LSTM on 20. At $h = 22$ the evidence is weaker, as expected with overlapping monthly targets and fewer effective observations: the inverse-MSE combination remains significantly better on 21 instruments and log-HAR on 17, ridge on the extended set on 10, and only the deep learners, the daily GARCH models and LightGBM on the HAR set are significantly worse than HAR on more than three instruments.

The model confidence sets tell the same story with the appropriate humility about multiplicity. With 27 models and roughly 2,400 observations per asset, the 90% MCS is large: at $h = 1$ it contains the inverse-MSE combination, the penalized linear models on the extended set, log-HAR, the pooled LightGBM and the random forest on the extended set for all 25 instruments, and the HAR benchmark itself for 19. The models that are eliminated most often are the daily GARCH models (in the set for 6 to 9 instruments), the two deep learners (5 each) and LightGBM on the HAR set (8). HARQ is retained for 21 instruments despite its poor mean loss, because the variance of its loss differential is as extreme as its mean. At $h = 22$ every model except the deep learners and LightGBM is in the set for at least three quarters of the instruments, which is the honest summary of the one-month horizon: differences in average loss are real but not large relative to their sampling variability.

5.5. Calibration and encompassing

Table 5 reports the median Mincer–Zarnowitz slope and R^2 across instruments, the share of instruments on which $(a,b) = (0,1)$ is rejected, and the encompassing tests against HAR, all at $h = 1$. The HAR forecast is well calibrated in the Mincer–Zarnowitz sense (median slope 1.03, rejection on 40 percent of the instruments) but explains less of the variation in next-day variance (median R^2

0.40) than the log-linear learners (0.44 to 0.46), the Realized GARCH (0.45) or the inverse-MSE combination (0.45). The log-linear models have slopes above one (1.09 for ridge on the extended set, 1.24 for log-HAR), the signature of forecasts that are too compressed: the lognormal and smearing retransformations recover the conditional mean on average but the forecasts under-react to the largest movements. Ridge on the extended set is nevertheless the model with the fewest rejections of unbiasedness (20 percent of the instruments). The tree ensembles and the deep learners are rejected on more than 90 percent of the instruments and explain far less of the target (median R^2 between 0.03 for the Transformer and 0.31 for XGBoost on the extended set), which identifies poor calibration, rather than a lack of information, as their main weakness. The encompassing regressions confirm that the information in the best learners is not contained in HAR: the coefficient on the rival forecast is significantly positive on 88 percent of the instruments for elastic net, 84 percent for ridge, LASSO and the pooled LightGBM, 80 percent for the Realized GARCH, and 76 percent for log-HAR and the combinations, but only on 36 percent for LightGBM and 20 to 24 percent for the HAR-set trees and the LSTM.

5.6. What the pooled model uses

Figure 6 reports the mean absolute SHAP values of the pooled LightGBM for the one-day target, computed on the temporal validation slice of the most recent 1,000-day window. The three HAR components account for the overwhelming majority of the attribution, with the weekly average (0.73 in log-variance units) more than twice as important as the daily (0.33) and monthly (0.32) averages. The remaining predictors contribute an order of magnitude less each: the day of the week (0.09, a weekend effect that only the cryptocurrencies can exhibit), the cross-sectional averages of log RV in the futures, crypto and FX classes (0.05 to 0.07), the negative semivariance and the log realized quarticity (0.06 each), the daily return (0.04) and the VIX (0.04); the jump share is last (0.003). The pooled learner therefore rediscovers the HAR cascade and adds a thin layer of cross-asset and calendar information, which is consistent with the small gap between it and log-HAR in Table 2.

6. Robustness

6.1. Sub-periods

Table 6 and Figure 7 report the one-day QLIKE ratio to HAR within four sub-periods: 2013–2019 (in practice December 2016 to December 2019, the start of the evaluation sample), the 2020 pandemic year, the 2021–2022 inflation and tightening episode, and January 2024 to July 2026. These ratios pool asset-days rather than averaging across instruments, so that the cryptocurrencies, which trade every day, weigh more than in Section 5 and the composition of the panel changes as the altcoins

enter. Two patterns emerge. The advantage of every log-linear model over HAR is largest in the early and pandemic sub-periods (ratios of 0.51 and 0.47 for ridge on the extended set, 0.54 and 0.50 for log-HAR, 0.55 and 0.51 for the Realized GARCH) and shrinks steadily thereafter (0.78 and 0.88 for ridge, 0.84 and 0.95 for log-HAR, 0.94 and 0.96 for the Realized GARCH in 2021–2022 and 2024–2026). The compression of the gains coincides with the maturing of the crypto markets and the disappearance of the extreme early-sample days discussed in Section 3, and with calmer conditions in the equity and FX markets; the HAR model in levels is at its worst when variance is most skewed. The second pattern concerns the trees: the asset-level LightGBM goes from a ratio of 0.65 before 2020 to 1.00 in 2024–2026, so that its relative performance decays faster than that of the linear learners, whereas the pooled LightGBM stays below one throughout (0.61, 0.48, 0.87, 0.93). The LSTM is worse than HAR in every sub-period (1.23 to 1.39), and the daily GARCH models are worse than HAR in every sub-period as well (1.13 to 1.68).

6.2. Stability over time

Figure 8 plots the Giacomini and Rossi (2010) fluctuation statistic for the one-day QLIKE differential between the asset-level LightGBM on the extended set and HAR, for one representative instrument per class, with a window of 30 percent of each evaluation sample and the two-sided 5% critical value of 3.01. Negative values favour the learner. For ETH the statistic is below the lower critical value on 52 percent of the dates, that is, the tree is significantly better than HAR for most of 2023 and 2024 and never significantly worse. For NVDA and GC the statistic is positive throughout and crosses the upper critical value on 8 and 14 percent of the dates respectively, so that HAR is locally significantly better than the tree in 2020–2022 and briefly in 2024 for NVDA, and in 2021–2022 and 2026 for GC. For GBPUSD the path never leaves the band. The fluctuation test thus confirms the class-specific reading of Section 5: the nonlinear learner adds value for the cryptocurrency and not for the equity, the future or the currency, and the relative performance is not constant within a class.

6.3. Transferability across assets

Table 7 addresses the question of whether volatility dynamics learned on other assets transfer to an asset the model has never seen. For one instrument per class we compare the HAR and LightGBM forecasts estimated on the asset's own history, the pooled LightGBM with the asset in the training pool, and the pooled LightGBM re-estimated with the asset excluded from training (leave-one-out) and applied out of pool. With the asset in the pool, pooling improves on the asset-level LightGBM for every representative at $h = 1$ (0.260 against 0.259 for ETH, 0.119 against 0.132 for GBPUSD, 0.130 against 0.153 for GC, 0.157 against 0.175 for NVDA) and matches or beats HAR for three of the four. Without the asset in the pool, transfer works only where the pool contains close substitutes:

for GBPUSD the out-of-pool forecast (0.123) is still better than HAR (0.128), because four other currency pairs with very similar dynamics remain in the pool; for GC the out-of-pool loss (0.154) is close to the asset-level LightGBM; for NVDA and ETH the out-of-pool losses (0.272 and 0.481) are far worse than HAR. At one month the degradation is severe for all four (0.175 to 2.99). The asset and class identifiers, and the asset-specific level and persistence they encode, are therefore essential ingredients of the pooled model; what transfers across assets is the shape of the HAR cascade, not a usable forecast for an unseen asset outside a homogeneous class.

6.4. Sampling frequency of the realized measure

Table 8 re-runs HAR, log-HAR and LightGBM on the extended set with the 1-minute RV and the realized kernel as both predictor and target, in place of the subsampled 5-minute RV, and reports the mean one-day QLIKE pooled over asset-days.[3] The ranking log-HAR $<$ LightGBM $<$ HAR holds for every measure. The level of the HAR loss, however, depends heavily on the measure (0.45 with 1-minute RV, 0.26 with the subsampled 5-minute RV, 0.24 with the kernel), whereas the log-HAR loss is almost invariant (0.19, 0.19, 0.20) and so is LightGBM's (0.22, 0.22, 0.23). The noisier the realized measure, the larger the apparent advantage of the log-linear and nonlinear models over HAR in levels: the ratio of log-HAR to HAR is 0.43 with 1-minute RV, 0.72 with 5-minute RV and 0.86 with the kernel. Part of what the learners "gain" is thus robustness to measurement noise in the predictors, which HAR in levels transmits linearly to the forecast—the attenuation mechanism that HARQ was designed to address and that, in our implementation, it addresses too aggressively.

6.5. Implausible crypto days

Table 9 recomputes the one-day crypto QLIKE ratios after removing the evaluation dates on which the realized target implies an annualized volatility above 500 percent, the data artefacts discussed in Section 3. The exclusion lowers every model's mean QLIKE, it leaves the ranking of the models unchanged, and it slightly widens the gap between the log-linear learners and HAR (ridge on the extended set moves from 0.55 to 0.49, log-HAR from 0.57 to 0.50, the pooled LightGBM from 0.59 to 0.48); the GARCH models remain 30 percent worse than HAR. The conclusions for the cryptocurrencies are therefore not driven by the extreme days, although the magnitude of every crypto loss is.

[3] This experiment was run by the author on 10 August 2026 with the HAR and LightGBM code paths, which have not changed since; it was not re-run for the present version, and the window-length sensitivity (500, 1,000 and 2,000 days) foreseen in the design has not been run at all. See Appendix C.

7. Economic value

7.1. Volatility timing

Table 10 reports, for the models carried into the economic-value exercise, the annualized Sharpe ratio and return of the volatility-timing strategy (23) net of 5 basis points, averaged over the 25 instruments, the average daily turnover, the Sharpe ratio at 0, 10 and 20 basis points, and the average utility gain (24) relative to HAR-based timing. Figure 9 plots the cumulative net return of four strategies on SPY. The first result is that statistical accuracy does not translate into timing performance. The Sharpe ratios of all strategies lie in a narrow band, from 0.26 for the asset-level LightGBM to 0.38 for the LSTM, with HAR at 0.34, and their ordering is almost the reverse of the QLIKE ordering: the two models with the highest net Sharpe ratios, the LSTM (0.38) and the GARCH(1,1) (0.35), are among the worst forecasters of Section 5, while the best forecasters—ridge on the extended set (0.30), the pooled LightGBM (0.29), LightGBM (0.26)—are at the bottom. The mechanism is turnover. The LSTM and the GARCH forecasts move slowly (average daily turnover of 0.013 and 0.033 in units of the risky weight) whereas the responsive forecasts trade three to nine times as much (0.054 for HAR, 0.085 for ridge, 0.118 for LightGBM), and at 5 basis points the cost of tracking next-day variance exceeds the benefit. Gross of costs the picture is different: at zero cost every strategy except the Transformer has a Sharpe ratio between 0.385 and 0.408 and the pooled LightGBM is the best (0.41), so that the accurate forecasters are at least as good as the smooth ones before costs. At 20 basis points the smooth forecasters dominate decisively (LSTM 0.33, GARCH 0.24, HAR 0.15, ridge 0.01, LightGBM -0.12).

The utility gains relative to HAR make the same point in annualized basis points: a mean-variance investor with $\gamma = 5$ would pay 33 basis points a year to switch from HAR to LSTM-based timing (median 40), 17 to switch to GARCH (median 10) and 10 to the equal-weight combination, but would require compensation of 47 basis points to switch to ridge on the extended set (median 24), 69 to the pooled LightGBM and 121 to the asset-level LightGBM. The gains are concentrated in the FX and futures markets, where the LSTM earns 89 and 54 basis points relative to HAR and the asset-level LightGBM loses 167 and 64; in equities the Realized GARCH (26), the combinations (14 to 23) and log-HAR (5) gain on HAR, ridge and GARCH are within 13 basis points of it, and the asset-level LightGBM loses 99. Two caveats apply. The timing exercise targets 10 percent annualized volatility using forecasts of *intraday* variance, so all strategies run above target and near the leverage cap on the low-volatility FX pairs; and the Sharpe ratios of the crypto and FX strategies are close to zero or negative for every model (Table 10 averages across classes), so that the cross-sectional average is driven by the positive drift of the equities and futures over 2017–2026. Within this design, however, the conclusion of Fleming et al. (2003) that realized-volatility forecasts have economic value does not extend to the *marginal* value of a more accurate forecaster once

transaction costs are charged: smoothness is worth more than accuracy to a volatility-targeting investor at realistic cost levels.

7.2. Value-at-Risk

Table 11 reports the mean violation rate of the one-day Gaussian VaR at the 1 and 5 percent levels and the share of instruments on which each model passes the Kupiec (1995) and Christoffersen (1998) tests at the 5% level. At the 5 percent level every forecaster is approximately calibrated: violation rates range from 4.9 percent (equal-weight combination) to 5.6 percent (EGARCH), the HAR model passes the unconditional-coverage test on 96 percent of the instruments and the conditional-coverage test on 84 percent, and the penalized linear models on the extended set pass both tests on 96 to 100 percent of the instruments—the best coverage record in the table, consistent with their superior QLIKE. The Transformer passes the conditional-coverage test on only 44 percent of the instruments and the LSTM on 56 percent, so that their violations cluster in time. At the 1 percent level all models fail: violation rates lie between 1.6 and 2.2 percent, the unconditional-coverage test is passed on at most 32 percent of the instruments and the conditional-coverage test on at most 40 percent (the combinations and the random forest). The common failure is the Gaussian quantile rather than the variance forecast; a Student-t or a filtered historical simulation tail would be required to use any of these forecasts for regulatory 1 percent VaR, and the results of Section 5 suggest that the choice of the variance forecaster would then matter less than the choice of the tail.

8. Conclusion

We have compared 27 forecasts of realized variance—six HAR specifications, four GARCH-type models, twelve penalized regressions and tree ensembles, two deep learners, a pooled multi-asset learner and two combinations—on 25 liquid instruments from four asset classes under one protocol, and pushed the comparison through to volatility timing and Value-at-Risk. Three conclusions follow.

First, the machine-learning gains over the HAR benchmark that the single-market literature reports do generalize across asset classes, but most of the gain is not machine learning. Forecasting the logarithm of realized variance with the HAR information set and a lognormal retransformation lowers the QLIKE by 25 percent at one day and 37 percent at one week relative to HAR in levels, on every class and on every instrument; penalized linear regressions on an extended set of eighteen predictors add about five points more at the short horizons and are the best individual forecasters, significantly better than HAR on 24 of 25 instruments and never significantly worse, with the inverse-MSE combination at their level. Tree ensembles and recurrent or attention networks, estimated asset by asset on roughly a thousand days, are worse than the linear learners everywhere and worse than

HAR at the monthly horizon; pooling across assets rescues the trees but does not make them better than log-HAR. These results side with Branco et al. (2024) rather than with the more optimistic reading of Christensen et al. (2023), and they suggest that the gains attributed to nonlinearity in the latter may partly reflect the choice of a benchmark in levels and the noise of the realized measure, since the advantage of every learner over HAR grows with the noise of the measure and shrinks in the calm, mature part of our sample.

Second, the gains are largest where volatility is most volatile. The QLIKE ratio of the best linear learner to HAR is 0.62 for the cryptocurrencies, 0.72 for equities, 0.78 for futures and 0.85 for FX at one day, and the same gradient appears in every log-linear model. Transfer to an asset absent from the training pool works only inside a homogeneous class (the currency pairs); elsewhere the asset identity is indispensable, so that what is common across assets is the HAR structure rather than a usable forecast.

Third, statistical accuracy and economic value diverge once transaction costs are charged. In volatility timing at 5 basis points the smoothest forecasters—the LSTM and the daily GARCH—have the highest net Sharpe ratios and positive utility gains over HAR, while the most accurate forecasters trade too much and lose 47 to 121 basis points a year relative to HAR. At the 5 percent VaR level every forecaster is calibrated and the linear learners have the best coverage record; at the 1 percent level all fail because of the Gaussian tail.

The study has limitations that the next version should address. The cryptocurrency bars are not cleaned and contain days on which even the subsampled 5-minute RV is implausible; removing them does not change the rankings but does change the level of every crypto loss, and a proper tick-level filter in the spirit of Brownlees and Gallo (2006) is the right remedy. The HARQ model behaves badly with a realized quarticity estimated from 78 to 288 five-minute returns, and a stricter filter or a jump-robust quarticity would be needed before its known gains on index data can be assessed here. The deep learners are re-estimated quarterly on single-asset samples and are given no pooled training, so that their poor showing is a statement about this design and not about sequence models in general. The sensitivity of the results to the estimation window (500 to 2,000 days) has not been run, and the sampling-frequency check was not re-run after the corrections made to the GARCH-family code in this version. Finally, the timing exercise targets volatility with intraday-variance forecasts and uses a single cost level for all markets; a class-specific cost schedule and close-to-close variance forecasts would sharpen the economic comparison. The replication package allows each of these extensions to be run from the same bars and under the same protocol.

References

Aït-Sahalia, Yacine, Per A. Mykland, and Lan Zhang, "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise," *Review of Financial Studies*, 2005, *18* (2), 351–416.

Alexander, Carol and Michael Dakos, "A Critical Investigation of Cryptocurrency Data and Analysis," *Quantitative Finance*, 2020, *20* (2), 173–188.

Andersen, Torben G. and Tim Bollerslev, "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review*, 1998, *39* (4), 885–905.

_ , _ , and Francis X. Diebold, "Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility," *Review of Economics and Statistics*, 2007, *89* (4), 701–720.

_ , _ , and Nour Meddahi, "Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities," *Econometrica*, 2005, *73* (1), 279–296.

_ , _ , Francis X. Diebold, and Heiko Ebens, "The Distribution of Realized Stock Return Volatility," *Journal of Financial Economics*, 2001, *61* (1), 43–76.

_ , _ , _ , and Paul Labys, "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 2001, *96* (453), 42–55.

_ , _ , _ , and _ , "Modeling and Forecasting Realized Volatility," *Econometrica*, 2003, *71* (2), 579–625.

Audrino, Francesco and Simon D. Knaus, "Lassoing the HAR Model: A Model Selection Perspective on Realized Volatility Dynamics," *Econometric Reviews*, 2016, *35* (8–10), 1485–1521.

Bandi, Federico M. and Jeffrey R. Russell, "Microstructure Noise, Realized Variance, and Optimal Sampling," *Review of Economic Studies*, 2008, *75* (2), 339–369.

Barndorff-Nielsen, Ole E. and Neil Shephard, "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society: Series B*, 2002, *64* (2), 253–280.

_ and _ , "Power and Bipower Variation with Stochastic Volatility and Jumps," *Journal of Financial Econometrics*, 2004, *2* (1), 1–37.

_ **and** _ , "Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation," *Journal of Financial Econometrics*, 2006, *4* (1), 1–30.

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

_ **,** _ **,** _ **, and** _ , "Realized Kernels in Practice: Trades and Quotes," *Econometrics Journal*, 2009, *12* (3), C1–C32.

_ **, Silja Kinnebrock, and Neil Shephard**, "Measuring Downside Risk: Realised Semivariance," in Tim Bollerslev, Jeffrey R. Russell, and Mark W. Watson, eds., *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, Oxford University Press, 2010, pp. 117–136.

Bates, John M. and Clive W. J. Granger, "The Combination of Forecasts," *Journal of the Operational Research Society*, 1969, *20* (4), 451–468.

Bollerslev, Tim, "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 1986, *31* (3), 307–327.

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

_ **, Benjamin Hood, John Huss, and Lasse Heje Pedersen**, "Risk Everywhere: Modeling and Managing Volatility," *Review of Financial Studies*, 2018, *31* (7), 2729–2773.

Branco, Rafael R., Alexandre Rubesam, and Mauricio Zevallos, "Forecasting Realized Volatility: Does Anything Beat Linear Models?," *Journal of Empirical Finance*, 2024, *78*, 101524.

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

Brownlees, Christian T. and Giampiero M. Gallo, "Financial Econometric Analysis at Ultra-High Frequency: Data Handling Concerns," *Computational Statistics & Data Analysis*, 2006, *51* (4), 2232–2245.

Buccheri, Giuseppe and Fulvio Corsi, "HARK the SHARK: Realized Volatility Modeling with Measurement Errors and Nonlinear Dependencies," *Journal of Financial Econometrics*, 2021, *19* (4), 614–649.

Bucci, Andrea, "Realized Volatility Forecasting with Neural Networks," *Journal of Financial Econometrics*, 2020, *18* (3), 502–531.

Catania, Leopoldo, Stefano Grassi, and Francesco Ravazzolo, "Forecasting Cryptocurrencies under Model and Parameter Instability," *International Journal of Forecasting*, 2019, *35* (2), 485–501.

Chen, Tianqi and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" 2016, pp. 785–794.

Christensen, Kim, Mathias Siggaard, and Bezirgen Veliyev, "A Machine Learning Approach to Volatility Forecasting," *Journal of Financial Econometrics*, 2023, *21* (5), 1680–1727.

Christoffersen, Peter F., "Evaluating Interval Forecasts," *International Economic Review*, 1998, *39* (4), 841–862.

Clements, Adam and Daniel P. A. Preve, "A Practical Guide to Harnessing the HAR Volatility Model," *Journal of Banking & Finance*, 2021, *133*, 106285.

Corsi, Fulvio, "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 2009, *7* (2), 174–196.

_ , Davide Pirino, and Roberto Renò, "Threshold Bipower Variation and the Impact of Jumps on Volatility Forecasting," *Journal of Econometrics*, 2010, *159* (2), 276–288.

Diebold, Francis X., "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests," *Journal of Business & Economic Statistics*, 2015, *33* (1), 1–9.

_ and Roberto S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 1995, *13* (3), 253–263.

Duan, Naihua, "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association*, 1983, *78* (383), 605–610.

Fleming, Jeff, Chris Kirby, and Barbara Ostdiek, "The Economic Value of Volatility Timing," *Journal of Finance*, 2001, *56* (1), 329–352.

_ , _ , and _ , "The Economic Value of Volatility Timing Using "Realized" Volatility," *Journal of Financial Economics*, 2003, *67* (3), 473–509.

Giacomini, Raffaella and Barbara Rossi, "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 2010, *25* (4), 595–620.

_ **and Halbert White**, "Tests of Conditional Predictive Ability," *Econometrica*, 2006, *74* (6), 1545–1578.

Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle, "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *Journal of Finance*, 1993, *48* (5), 1779–1801.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 2020, *33* (5), 2223–2273.

Hansen, Peter R. and Asger Lunde, "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?," *Journal of Applied Econometrics*, 2005, *20* (7), 873–889.

_ **and** _ , "Consistent Ranking of Volatility Models," *Journal of Econometrics*, 2006, *131* (1–2), 97–121.

_ **and** _ , "Realized Variance and Market Microstructure Noise," *Journal of Business & Economic Statistics*, 2006, *24* (2), 127–161.

_ **,** _ **, and James M. Nason**, "The Model Confidence Set," *Econometrica*, 2011, *79* (2), 453–497.

Hansen, Peter Reinhard and Zhuo Huang, "Exponential GARCH Modeling with Realized Measures of Volatility," *Journal of Business & Economic Statistics*, 2016, *34* (2), 269–287.

_ **,** _ **, and Howard Howan Shek**, "Realized GARCH: A Joint Model for Returns and Realized Measures of Volatility," *Journal of Applied Econometrics*, 2012, *27* (6), 877–906.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold, "Tests for Forecast Encompassing," *Journal of Business & Economic Statistics*, 1998, *16* (2), 254–259.

Hochreiter, Sepp and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997, *9* (8), 1735–1780.

Katsiampa, Paraskevi, "Volatility Estimation for Bitcoin: A Comparison of GARCH Models," *Economics Letters*, 2017, *158*, 3–6.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in "Advances in Neural Information Processing Systems," Vol. 30 2017, pp. 3146–3154.

Kupiec, Paul H., "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, 1995, *3* (2), 73–84.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard, "Does Anything Beat 5-Minute RV? A Comparison of Realized Measures Across Multiple Asset Classes," *Journal of Econometrics*, 2015, *187* (1), 293–311.

Lundberg, Scott M. and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," in "Advances in Neural Information Processing Systems," Vol. 30 2017, pp. 4765–4774.

Mincer, Jacob A. and Victor Zarnowitz, "The Evaluation of Economic Forecasts," in Jacob A. Mincer, ed., *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, New York: National Bureau of Economic Research, 1969, pp. 3–46.

Müller, Ulrich A., Michel M. Dacorogna, Rakhal D. Davé, Richard B. Olsen, Olivier V. Pictet, and Jacob E. von Weizsäcker, "Volatilities of Different Time Resolutions — Analyzing the Dynamics of Market Components," *Journal of Empirical Finance*, 1997, *4* (2–3), 213–239.

Nelson, Daniel B., "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 1991, *59* (2), 347–370.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 1987, *55* (3), 703–708.

Patton, Andrew J., "Volatility Forecast Comparison Using Imperfect Volatility Proxies," *Journal of Econometrics*, 2011, *160* (1), 246–256.

_ , "Comparing Possibly Misspecified Forecasts," *Journal of Business & Economic Statistics*, 2020, *38* (4), 796–809.

_ **and Kevin Sheppard**, "Evaluating Volatility and Correlation Forecasts," in Torben G. Andersen, Richard A. Davis, Jens-Peter Kreiß, and Thomas Mikosch, eds., *Handbook of Financial Time Series*, Berlin: Springer, 2009, pp. 801–838.

_ **and** _ , "Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility," *Review of Economics and Statistics*, 2015, *97* (3), 683–697.

Politis, Dimitris N. and Joseph P. Romano, "The Stationary Bootstrap," *Journal of the American Statistical Association*, 1994, *89* (428), 1303–1313.

Rahimikia, Eghbal and Ser-Huang Poon, "Machine Learning for Realised Volatility Forecasting," Technical Report, SSRN Working Paper No. 3707796 2020.

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

Timmermann, Allan, "Forecast Combinations," in Graham Elliott, Clive W. J. Granger, and Allan Timmermann, eds., *Handbook of Economic Forecasting*, Vol. 1, Elsevier, 2006, pp. 135–196.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in "Advances in Neural Information Processing Systems," Vol. 30 2017, pp. 5998–6008.

West, Kenneth D., "Asymptotic Inference about Predictive Ability," *Econometrica*, 1996, *64* (5), 1067–1084.

Zhang, Chao, Yihuang Zhang, Mihai Cucuringu, and Zhongmin Qian, "Volatility Forecasting with Machine Learning and Intraday Commonality," *Journal of Financial Econometrics*, 2024, *22* (2), 492–530.

Zhang, Lan, Per A. Mykland, and Yacine Aït-Sahalia, "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 2005, *100* (472), 1394–1411.

Zou, Hui and Trevor Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B*, 2005, *67* (2), 301–320.

A. Asset-level results

Table 12 reports the mean one-day QLIKE of the headline models for each of the 25 instruments on the common evaluation sample; the lowest loss in each row is in bold. The asset-level Diebold–Mariano statistics, MCS p-values, Mincer–Zarnowitz and encompassing regressions, timing statistics and VaR backtests are available as machine-readable files in `results/reproduced/` of the replication package.

B. Hyperparameter grids and implementation details

Hyperparameters of the tabular learners are selected on the last 20% of each estimation window, in time order, from the following grids, and re-selected once a year: ridge $\lambda \in \{0.1, 1, 10\}$; LASSO $\lambda \in \{0.0005, 0.005, 0.05\}$; elastic net $\lambda \in \{0.001, 0.01\}$ with ℓ_1 share $\in \{0.3, 0.7\}$; random forest (300 trees) with minimum leaf size $\in \{5, 20\}$ and feature fraction $\in \{1/3, \sqrt{p}/p\}$; XGBoost (400 trees, learning rate 0.05, subsampling 0.8) with maximum depth $\in \{3, 5\}$; LightGBM (400 trees, learning rate 0.05, subsampling 0.8) with $\{15, 31\}$ leaves. Penalized regressions act on standardized predictors. The pooled LightGBM uses 500 trees, learning rate 0.05, 63 leaves and a minimum of 50

observations per leaf, with the ticker and asset-class identifiers as categorical features. The LSTM and the Transformer are trained for at most 50 epochs with Adam (learning rate 10^{-3}, batches of 128 sequences) and early stopping with patience 8 on the validation MSE of the three log targets. All random seeds are fixed in `config.yaml`.

C. Reproducibility

The replication package (`scripts/01` to `scripts/08`) downloads the bars, builds the realized measures, produces the forecasts, runs the evaluation, robustness and economic-value blocks, and generates every figure and table of the paper; no number in the text is typed by hand. On an 18-core Apple M5 Max the econometric family takes about six minutes for the 25 assets, the tabular learners about two hours, the deep learners about twenty minutes, and the pooled and leave-one-out LightGBM models about fifteen minutes; the evaluation blocks run in a few minutes. Section 6 lists the experiments of the design that were not executed for this version.

Table 1. Descriptive statistics of daily realized volatility by instrument. Volatility columns annualize the subsampled 5-minute RV ($\sqrt{252 \cdot RV}$, in percent). Skewness and first-order autocorrelation refer to $\ln RV$. "Jump days" is the share of days with a significant BNS jump statistic at $\alpha = 0.1\%$. Sample: January 2013 (or first available date) to July 2026.

Ticker	Days	Mean vol. (%)	Median vol. (%)	P95 vol. (%)	Skew. lnRV	AC(1) lnRV	Jump days (%)
Equities/ETFs							
AAPL	3,413	18.8	16.6	34.2	0.55	0.73	4.6
GLD	3,396	8.8	7.8	16.4	0.65	0.59	3.9
IWM	3,405	15.3	13.3	28.5	0.77	0.77	2.5
JPM	3,388	17.6	15.6	30.7	0.98	0.74	4.6
MSFT	3,397	17.8	15.9	32.0	0.65	0.75	4.7
NVDA	3,401	30.3	26.4	57.0	0.60	0.75	4.4
QQQ	3,410	13.4	11.3	27.2	0.47	0.78	2.8
SPY	3,414	10.2	8.4	21.3	0.53	0.79	2.8
TLT	3,394	8.5	7.7	15.0	0.73	0.70	2.8
XOM	3,389	18.0	15.7	33.6	0.59	0.84	4.3
FX							
AUDUSD	3,521	9.6	8.9	16.1	0.41	0.68	8.9
EURUSD	3,522	7.2	6.6	12.8	0.30	0.69	7.7
GBPUSD	3,521	8.1	7.4	13.5	0.65	0.70	7.7
USDCAD	3,521	6.7	6.2	11.4	0.24	0.69	6.3
USDJPY	3,522	7.9	7.0	14.7	0.24	0.72	10.2
Crypto							
ADA	2,750	106.8	81.4	221.7	-2.75	0.65	25.2
BTC	4,498	97.8	50.3	372.7	0.95	0.90	11.8
ETH	3,528	108.4	75.2	289.7	0.46	0.88	12.1
SOL	1,917	81.4	69.7	162.2	0.53	0.76	5.9
XRP	3,373	137.3	77.2	478.6	0.85	0.90	18.3
Futures							
CL	3,502	33.4	28.3	64.7	0.78	0.87	10.5
ES	3,501	13.4	10.9	28.3	0.60	0.81	6.6
GC	3,501	14.1	12.5	25.1	0.71	0.70	8.9
NQ	3,498	16.9	14.2	35.2	0.38	0.80	6.3
ZN	3,216	5.2	4.6	9.3	0.82	0.68	19.5

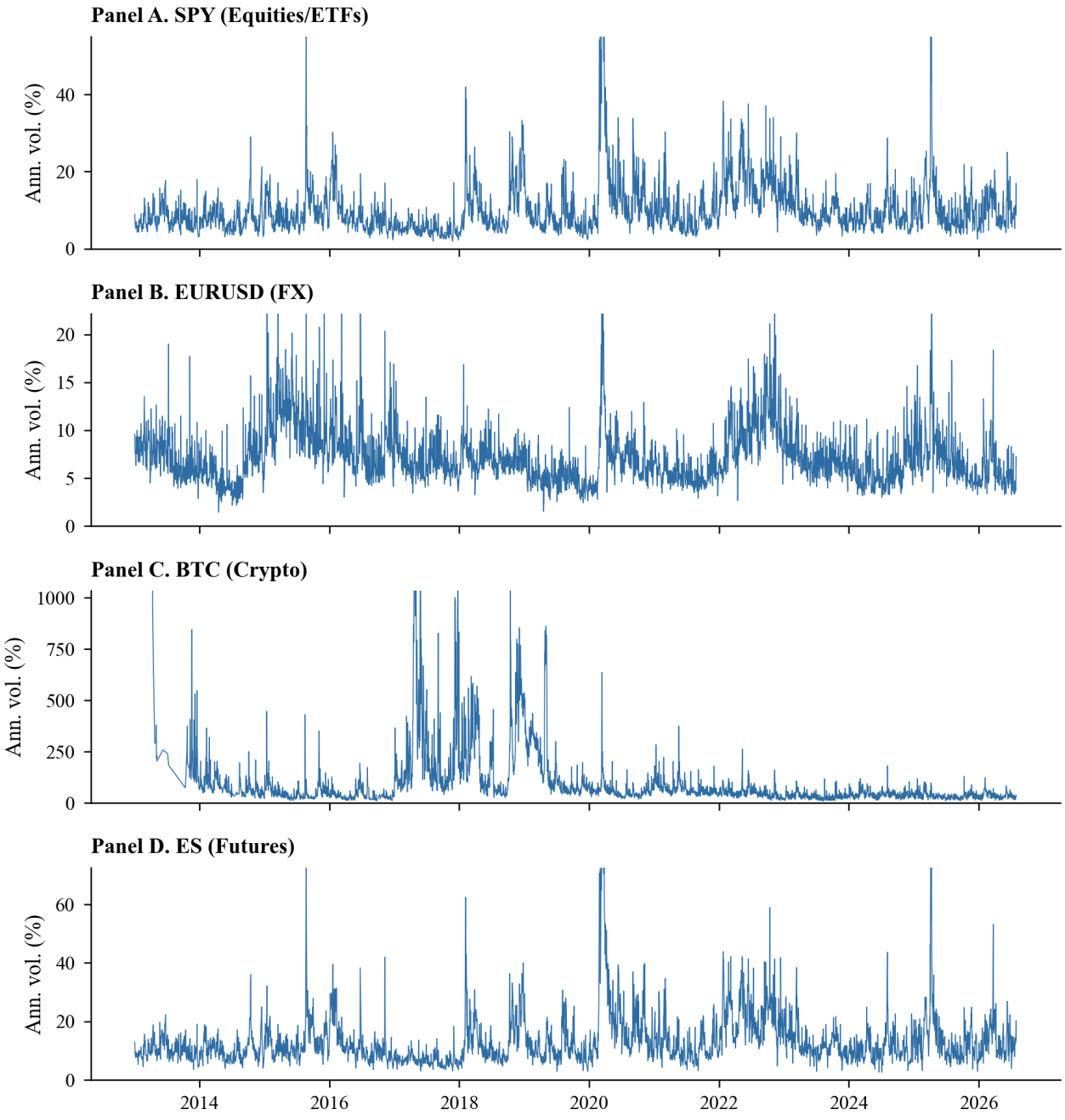


Figure 1. Annualized realized volatility (subsampled 5-minute RV), one representative instrument per asset class, 2013–2026. The vertical axis of each panel is capped at the 99.5th percentile of the series.

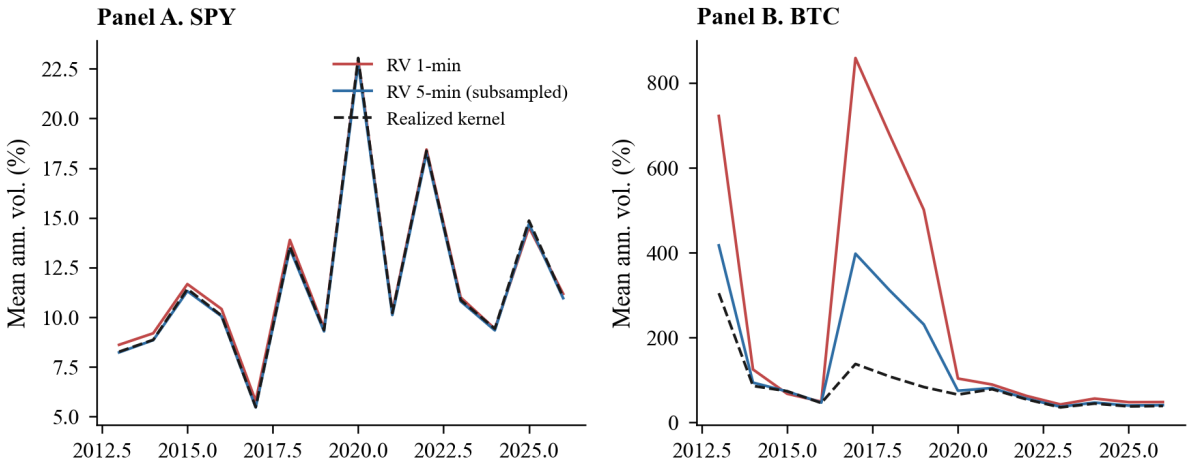


Figure 2. Yearly mean annualized volatility by realized measure. The divergence between the 1-minute RV and the noise-robust measures for early-sample BTC reflects market microstructure noise.

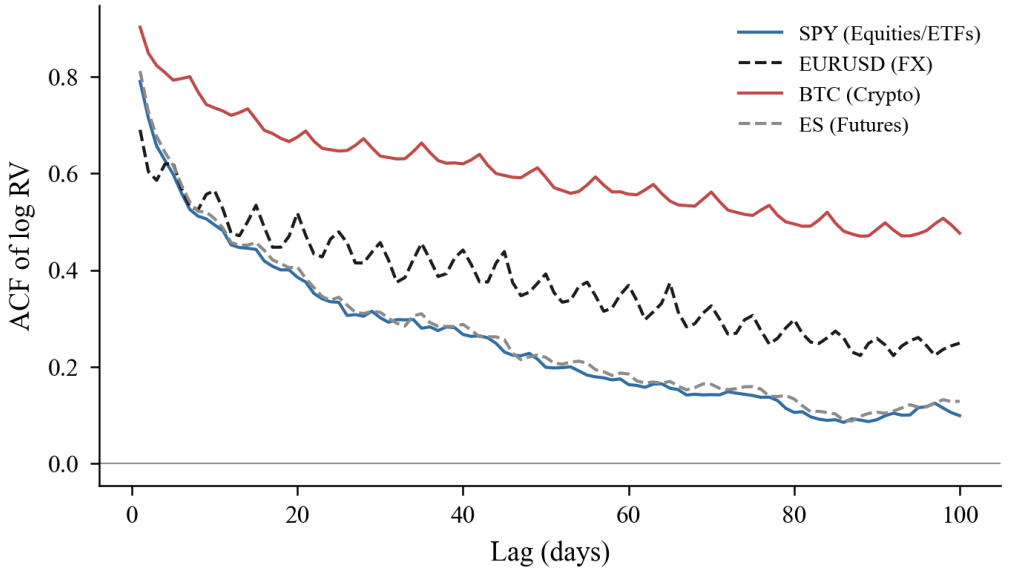


Figure 3. Autocorrelation functions of daily $\ln RV$ (subsampled 5-minute measure), representative instruments.

Table 2. Out-of-sample QLIKE by model and horizon. Mean QLIKE across the 25 instruments (each instrument's mean loss computed on its common evaluation sample, then averaged) and ratio to the HAR benchmark; ratios below one favour the model. The lowest value in each column is in bold. Suffixes -H and -X denote the HAR and extended information sets of the tabular learners.

Model	QLIKE h=1	Ratio h=1	QLIKE h=5	Ratio h=5	QLIKE h=22	Ratio h=22
A. GARCH family						
GARCH	0.338	1.319	0.302	1.117	0.341	1.184
GJR	0.338	1.321	0.302	1.118	0.340	1.182
EGARCH	0.401	1.566	0.335	1.238	0.355	1.234
RealGARCH	0.200	0.782	0.179	0.661	0.253	0.879
B. HAR family						
HAR	0.256	1.000	0.271	1.000	0.288	1.000
HAR-J	0.241	0.942	0.238	0.879	0.291	1.010
HAR-CJ	0.240	0.936	0.238	0.878	0.304	1.057
SHAR	0.253	0.988	0.262	0.968	0.289	1.003
HARQ	0.516	2.015	0.466	1.722	0.482	1.675
LogHAR	0.193	0.752	0.169	0.625	0.227	0.790
C. Machine learning — HAR information set						
Ridge-H	0.191	0.746	0.169	0.626	0.230	0.797
LASSO-H	0.193	0.752	0.172	0.637	0.236	0.819
ElasticNet-H	0.191	0.747	0.170	0.627	0.230	0.799
RF-H	0.218	0.851	0.204	0.753	0.276	0.960
XGBoost-H	0.232	0.904	0.219	0.809	0.318	1.106
LightGBM-H	0.239	0.935	0.224	0.828	0.320	1.110
D. Machine learning — extended information set						
Ridge-X	**0.180**	**0.702**	**0.164**	**0.606**	0.227	0.789
LASSO-X	0.182	0.710	0.166	0.613	0.231	0.800
ElasticNet-X	0.180	0.704	0.164	0.607	0.226	0.786
RF-X	0.201	0.786	0.194	0.716	0.273	0.947
XGBoost-X	0.205	0.800	0.217	0.802	0.323	1.121
LightGBM-X	0.218	0.853	0.230	0.849	0.333	1.156
E. Deep learning and pooled						
LSTM	0.340	1.329	0.283	1.046	0.303	1.053
Transformer	0.484	1.889	0.399	1.477	0.420	1.459
Pooled-LGBM	0.196	0.765	0.189	0.700	0.275	0.954
F. Forecast combinations						
Comb-Mean	0.204	0.796	0.185	0.684	0.250	0.867
Comb-InvMSE	0.185	0.723	0.165	0.610	**0.223**	**0.774**

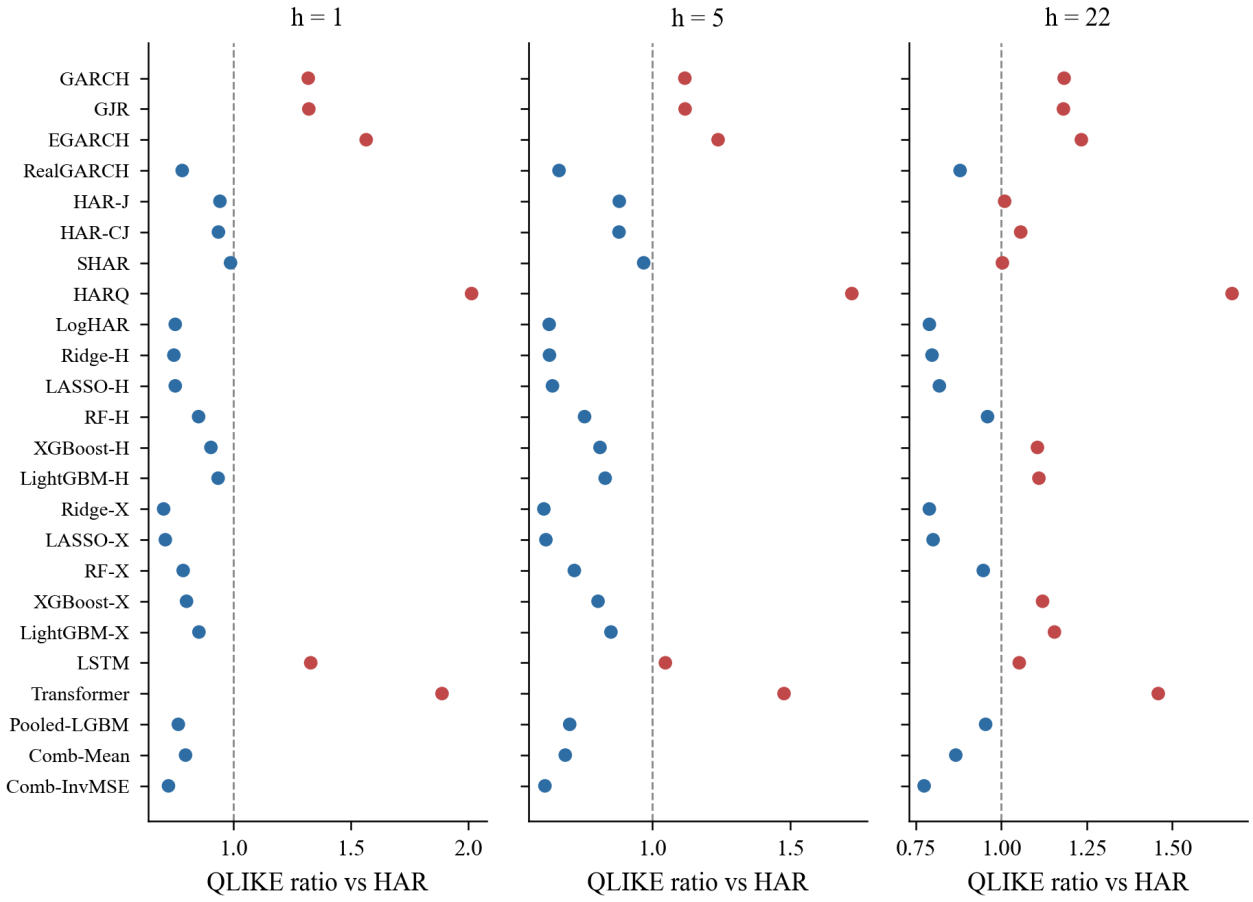


Figure 4. QLIKE ratio to HAR by model and horizon. Blue markers denote models with a lower mean QLIKE than HAR, red markers a higher one. Source: `results/reproduced/losses_overall.csv`.

Table 3. QLIKE ratio to HAR by asset class, $h = 1$ and $h = 22$. Mean over the instruments of each class of the instrument-level mean QLIKE, divided by the same quantity for HAR.

	$h = 1$				$h = 22$			
Model	Equities	FX	Crypto	Futures	Equities	FX	Crypto	Futures
A. GARCH family								
GARCH	1.067	1.238	1.597	1.170	0.920	1.169	1.641	0.914
GJR	1.049	1.322	1.598	1.152	0.894	1.269	1.642	0.904
EGARCH	1.442	1.808	1.650	1.394	0.979	1.345	1.614	1.020
RealGARCH	0.766	0.925	0.709	0.907	0.883	0.912	0.832	0.936
B. HAR family								
HAR-J	1.030	1.070	0.816	0.993	1.001	1.000	1.029	0.999
HAR-CJ	1.038	0.978	0.818	1.003	1.023	0.976	1.153	0.995
SHAR	1.013	1.027	0.951	1.001	1.010	0.987	0.998	1.006
HARQ	1.755	1.613	1.916	3.286	1.586	0.995	1.608	2.242
LogHAR	0.763	0.912	0.652	0.879	0.837	0.911	0.670	0.847
C. Machine learning — HAR information set								
Ridge-H	0.761	0.914	0.638	0.879	0.842	0.912	0.685	0.851
LASSO-H	0.766	0.920	0.639	0.903	0.856	0.923	0.698	0.905
ElasticNet-H	0.762	0.915	0.638	0.882	0.844	0.912	0.686	0.856
RF-H	0.930	0.968	0.678	1.071	0.992	0.979	0.774	1.199
XGBoost-H	0.948	1.055	0.748	1.126	1.150	1.141	0.883	1.376
LightGBM-H	1.013	1.111	0.739	1.168	1.154	1.183	0.889	1.362
D. Machine learning — extended information set								
Ridge-X	0.715	**0.852**	**0.621**	**0.778**	0.856	0.977	0.647	**0.823**
LASSO-X	0.723	0.860	0.629	0.783	0.849	0.970	0.670	0.854
ElasticNet-X	**0.715**	0.856	0.623	0.780	0.849	0.963	0.647	0.826
RF-X	0.839	0.921	0.624	1.014	0.993	0.954	0.749	1.188
XGBoost-X	0.877	0.972	0.630	0.965	1.201	1.135	0.892	1.347
LightGBM-X	0.947	1.019	0.668	1.028	1.225	1.185	0.937	1.377
E. Deep learning and pooled								
LSTM	1.451	2.006	0.883	1.752	1.088	1.557	0.822	1.148
Transformer	2.022	1.783	1.403	3.099	1.469	1.769	1.086	1.920
Pooled-LGBM	0.794	0.943	0.648	0.883	1.071	1.033	0.773	1.007
F. Forecast combinations								
Comb-Mean	0.770	0.974	0.740	0.864	0.862	0.940	0.836	0.894
Comb-InvMSE	0.740	0.889	0.625	0.824	**0.827**	**0.880**	**0.646**	0.843

Table 4. Diebold–Mariano tests against HAR and 90% model confidence sets. Share of the 25 instruments (in percent) on which the model's QLIKE is significantly lower ("better") or higher ("worse") than that of HAR at the 5% level (DM test with Bartlett bandwidth $h-1$), and share of instruments for which the model belongs to the 90% MCS (stationary bootstrap, 5,000 replications).

	Sig. better than HAR (%)			Sig. worse than HAR (%)			In 90% MCS (%)		
Model	h=1	h=5	h=22	h=1	h=5	h=22	h=1	h=5	h=22
A. GARCH family									
GARCH	4	4	8	80	64	20	24	48	80
GJR	4	4	8	76	52	20	36	60	76
EGARCH	0	0	0	92	64	24	24	52	76
RealGARCH	68	68	20	0	0	0	92	96	96
B. HAR family									
HAR	—	—	—	—	—	—	76	72	88
HAR-J	12	8	0	4	12	8	72	72	88
HAR-CJ	20	16	12	36	16	4	72	76	88
SHAR	8	20	8	28	20	8	72	72	88
HARQ	4	8	12	20	4	0	84	84	88
LogHAR	84	88	68	0	0	0	100	100	100
C. Machine learning — HAR information set									
Ridge-H	88	88	64	0	0	0	92	100	100
LASSO-H	84	84	60	0	0	0	96	100	100
ElasticNet-H	88	88	64	0	0	0	92	100	100
RF-H	20	24	16	24	8	4	88	100	100
XGBoost-H	8	8	12	48	32	8	48	68	80
LightGBM-H	8	8	12	64	48	20	32	32	52
D. Machine learning — extended information set									
Ridge-X	96	88	40	0	0	0	100	100	100
LASSO-X	88	84	32	0	0	0	100	100	100
ElasticNet-X	96	88	36	0	0	0	100	100	100
RF-X	44	28	12	4	4	0	100	100	100
XGBoost-X	24	8	8	16	8	8	92	92	84
LightGBM-X	16	8	4	32	36	12	68	80	56
E. Deep learning and pooled									
LSTM	4	8	8	80	84	40	20	36	52
Transformer	0	4	8	88	88	60	20	28	40
Pooled-LGBM	60	16	16	0	0	4	100	100	96
F. Forecast combinations									
Comb-Mean	76	76	60	4	0	0	92	96	88
Comb-InvMSE	96	88	84	0	0	0	100	100	100

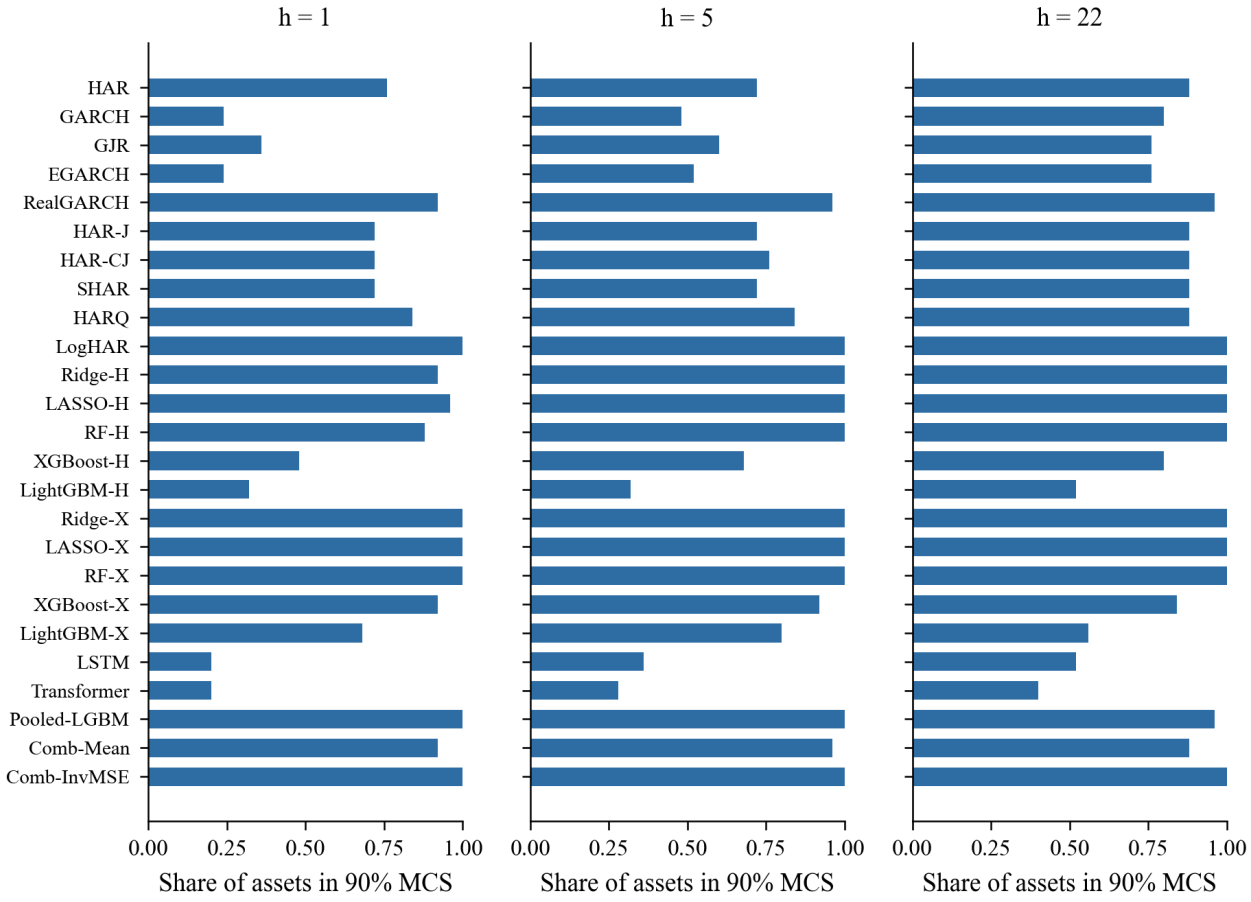


Figure 5. Share of the 25 instruments for which each model belongs to the 90% model confidence set, by horizon. Source: `results/reproduced/mcs_inclusion.csv`.

Table 5. Calibration and encompassing, $h = 1$. Median across the 25 instruments of the Mincer–Zarnowitz slope $\hat{b}$ and R^2 of regression (19); share of instruments on which the joint null $(a, b) = (0, 1)$ is rejected at 5% (HAC Wald test); median coefficient $\hat{b}_2$ on the rival forecast in the encompassing regression (20) and share of instruments on which b_2 is significantly positive at 5% ("adds information").

	Mincer–Zarnowitz			Encompassing (vs HAR)	
Model	Med. $\hat{\beta}$	Med. R^2	Rej. (%)	Med. $\hat{b}_2$	Adds info (%)
A. GARCH family					
GARCH	1.08	0.29	64	0.39	60
GJR	1.04	0.29	64	0.32	60
EGARCH	0.54	0.10	92	0.13	64
RealGARCH	1.29	0.45	52	0.98	80
B. HAR family					
HAR	1.03	0.40	40	—	—
HAR-J	1.01	0.40	40	0.41	24
HAR-CJ	1.01	0.39	36	0.35	28
SHAR	1.02	0.39	44	0.39	40
HARQ	0.94	0.38	32	0.32	32
LogHAR	1.24	0.44	36	1.10	76
C. Machine learning — HAR information set					
Ridge-H	1.20	0.44	28	1.08	72
LASSO-H	1.25	0.44	48	1.09	60
ElasticNet-H	1.22	0.44	28	1.08	72
RF-H	1.14	0.23	60	0.06	20
XGBoost-H	0.87	0.20	92	0.00	20
LightGBM-H	0.87	0.21	92	0.02	24
D. Machine learning — extended information set					
Ridge-X	1.09	0.46	20	0.96	84
LASSO-X	1.10	0.46	40	1.01	84
ElasticNet-X	1.13	0.46	28	1.01	88
RF-X	1.32	0.29	92	0.38	40
XGBoost-X	1.15	0.31	92	0.40	40
LightGBM-X	1.17	0.27	96	0.38	36
E. Deep learning and pooled					
LSTM	0.86	0.16	92	0.12	24
Transformer	0.90	0.03	92	-0.05	4
Pooled-LGBM	1.05	0.44	84	0.76	84
F. Forecast combinations					
Comb-Mean	1.27	0.42	88	1.05	76
Comb-InvMSE	1.22	0.45	56	1.49	76

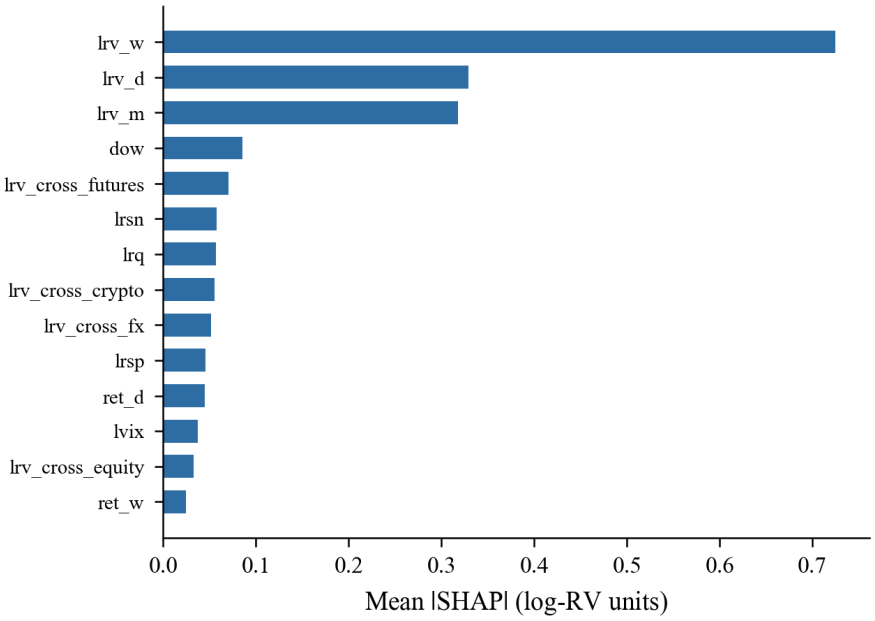


Figure 6. Mean absolute SHAP value by predictor, pooled LightGBM, $h = 1$ (log-variance units). Source: `results/reproduced/shap_h1.csv`.

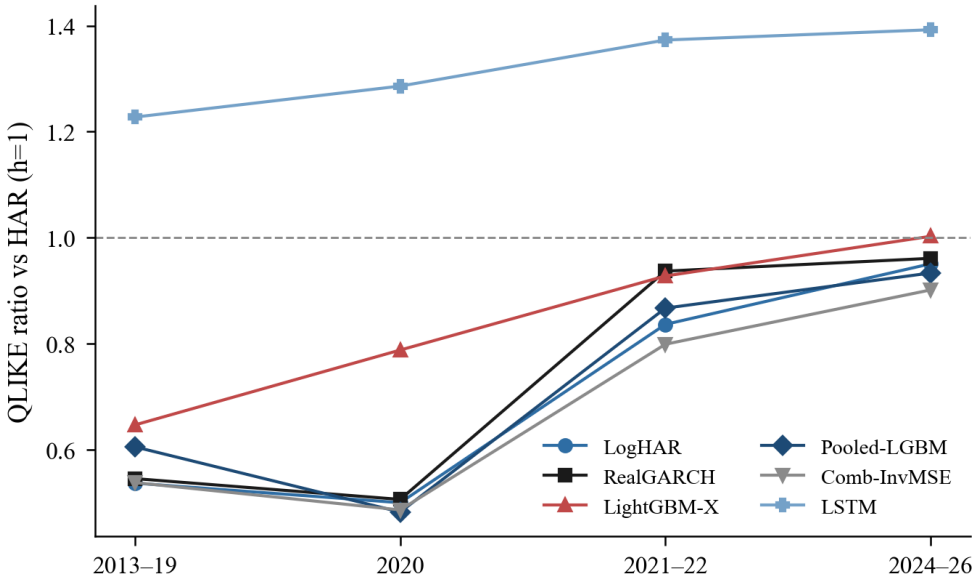


Figure 7. QLIKE ratio to HAR across sub-periods, $h = 1$, selected models. Source: `results/reproduced/subperiod_losses.csv`.

Table 6. QLIKE ratio to HAR by sub-period, $h = 1$. Asset-days pooled within each sub-period. The first sub-period starts with the evaluation sample in December 2016.

Model	2013–19	2020	2021–22	2024–26
A. GARCH family				
GARCH	1.468	1.126	1.678	1.169
GJR	1.447	1.235	1.683	1.140
EGARCH	1.733	1.464	1.679	1.547
RealGARCH	0.546	0.506	0.937	0.961
B. HAR family				
HAR	1.000	1.000	1.000	1.000
HAR-J	0.648	1.001	1.059	1.026
HAR-CJ	0.638	1.021	1.032	1.015
SHAR	0.645	1.214	1.173	1.007
HARQ	2.630	3.326	1.479	1.670
LogHAR	0.537	0.500	0.837	0.951
C. Machine learning — HAR information set				
Ridge-H	0.541	0.497	0.817	0.945
LASSO-H	0.544	0.512	0.823	0.950
ElasticNet-H	0.541	0.498	0.818	0.946
RF-H	0.613	0.888	0.840	1.023
XGBoost-H	0.619	0.848	0.989	1.090
LightGBM-H	0.665	0.879	1.000	1.111
D. Machine learning — extended information set				
Ridge-X	**0.510**	**0.467**	**0.775**	**0.883**
LASSO-X	0.517	0.469	0.785	0.892
ElasticNet-X	0.511	0.467	0.777	0.885
RF-X	0.550	0.802	0.804	0.936
XGBoost-X	0.592	0.723	0.860	0.955
LightGBM-X	0.647	0.788	0.928	1.003
E. Deep learning and pooled				
LSTM	1.228	1.286	1.373	1.392
Transformer	1.833	3.202	1.718	1.567
Pooled-LGBM	0.606	0.483	0.867	0.933
F. Forecast combinations				
Comb-Mean	0.607	0.608	0.899	0.944
Comb-InvMSE	0.538	0.487	0.799	0.902

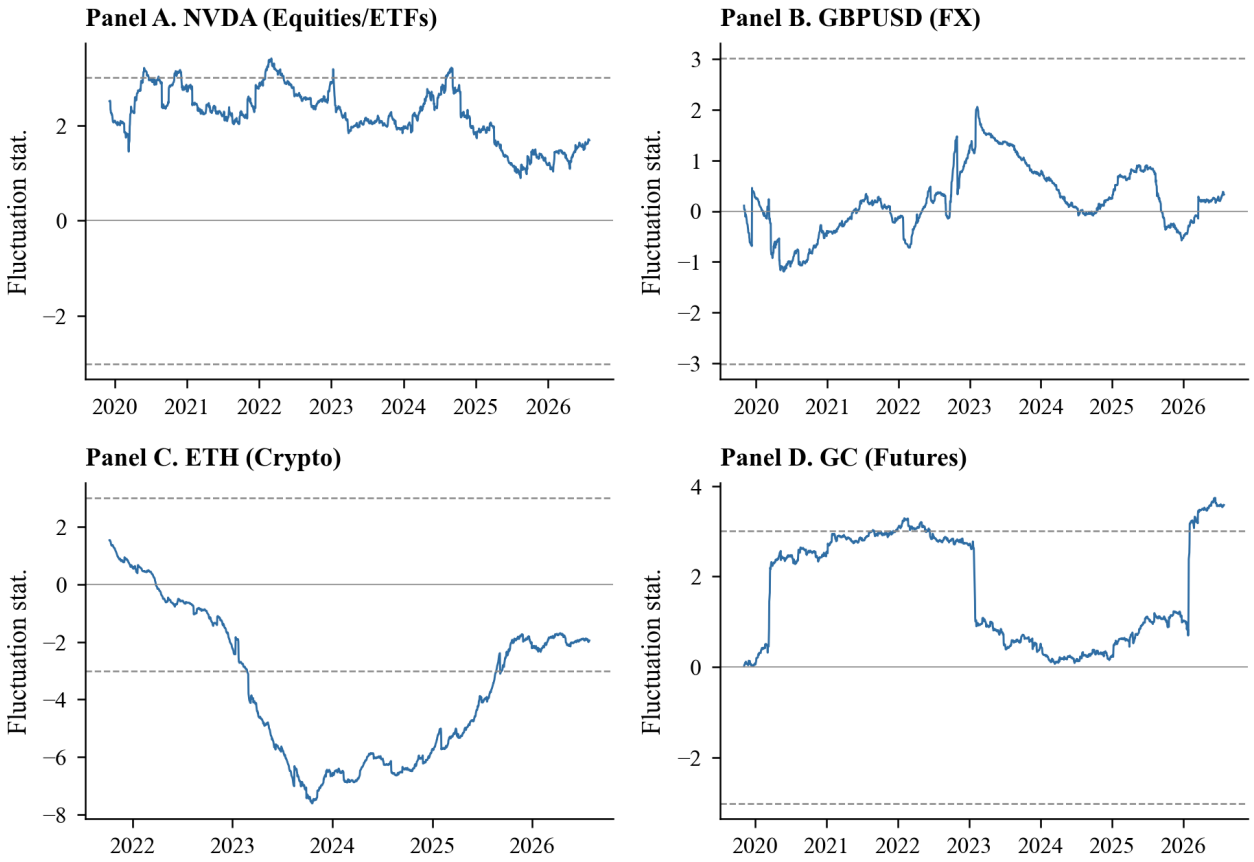


Figure 8. Giacomini–Rossi fluctuation statistic, LightGBM (extended set) versus HAR, $h = 1$, window of 30% of the evaluation sample. Dashed lines are the two-sided 5% critical values (± 3.01); values below the lower line indicate that LightGBM is significantly better locally. Source: `results/reproduced/gr_fluctuation.csv`.

Table 7. Cross-asset transferability, mean QLIKE at $h = 1$, one representative instrument per class. "Asset in pool" is the pooled LightGBM of Table 2; "asset excluded" is the same model re-estimated without the instrument in the training pool and applied to it out of pool. Common sample of the four forecasts per instrument.

Model	NVDA	GBPUSD	ETH	GC
HAR (own asset)	0.155	0.128	0.317	0.131
LightGBM (own asset)	0.175	0.132	0.259	0.153
Pooled LightGBM (asset in pool)	0.157	0.119	0.260	0.130
Pooled LightGBM (asset excluded)	0.272	0.123	0.481	0.154

Table 8. Sensitivity to the realized measure, mean QLIKE at $h = 1$ pooled over asset-days, with the ratio to HAR within each measure in parentheses. The measure is used both as predictor and as target.

Model	1-min RV	5-min ss RV	Realized kernel
HAR	0.451 (1.00)	0.263 (1.00)	0.238 (1.00)
LogHAR	0.192 (0.42)	0.189 (0.72)	0.204 (0.86)
LightGBM	0.218 (0.48)	0.218 (0.83)	0.235 (0.99)

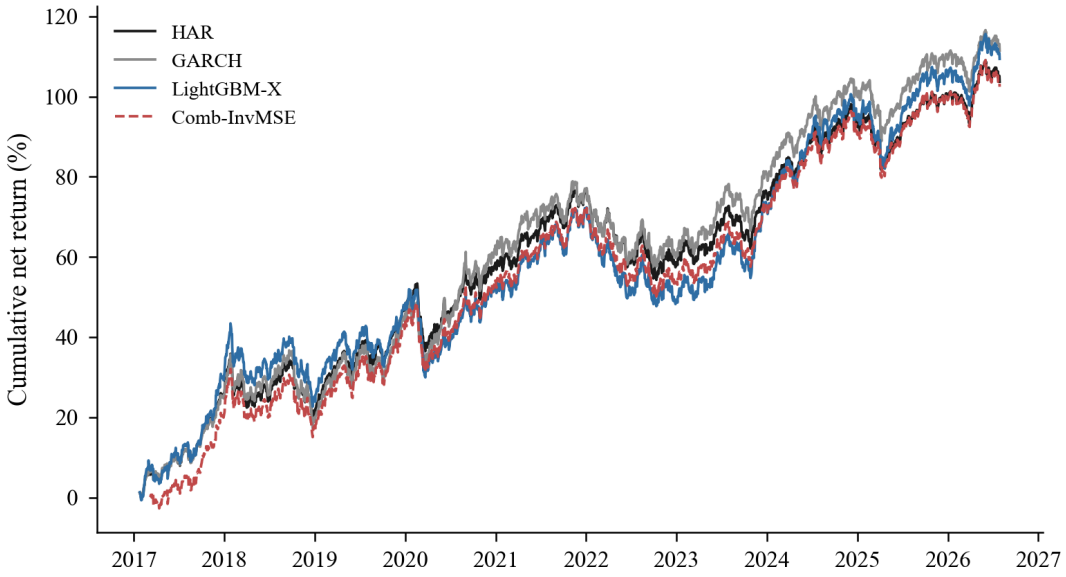


Figure 9. Cumulative net return (5 basis points) of volatility timing on SPY under four forecasters, 2017–2026. Source: `data/processed/merged_forecasts.parquet` and `src/wp12/backtest.py`.

Table 9. Crypto results with and without implausible days, $h = 1$. Mean QLIKE pooled over crypto asset-days and ratio to HAR, on the full common sample and after removing dates whose realized target exceeds 500% annualized volatility.

	All crypto days ($N = 10,951$)		Excl. days >500% ann. ($N = 10,802$)	
Model	QLIKE	Ratio	QLIKE	Ratio
A. GARCH family				
GARCH	0.992	1.671	0.710	1.303
GJR	0.992	1.670	0.710	1.303
EGARCH	1.022	1.722	0.741	1.361
RealGARCH	0.367	0.618	0.299	0.549
B. HAR family				
HAR	0.594	1.000	0.545	1.000
HAR-J	0.440	0.742	0.390	0.716
HAR-CJ	0.441	0.743	0.393	0.721
SHAR	0.539	0.908	0.412	0.757
HARQ	1.323	2.229	1.267	2.326
LogHAR	0.341	0.574	0.272	0.499
C. Machine learning — HAR information set				
Ridge-H	0.335	0.564	0.273	0.502
LASSO-H	0.335	0.565	0.274	0.503
ElasticNet-H	0.335	0.564	0.274	0.502
RF-H	0.364	0.612	0.275	0.506
XGBoost-H	0.388	0.654	0.309	0.568
LightGBM-H	0.389	0.656	0.311	0.570
D. Machine learning — extended information set				
Ridge-X	0.326	0.549	0.268	0.491
LASSO-X	0.330	0.556	0.271	0.497
ElasticNet-X	0.327	0.551	0.268	0.492
RF-X	0.330	0.556	0.252	0.462
XGBoost-X	0.334	0.563	0.255	0.467
LightGBM-X	0.361	0.608	0.271	0.498
E. Deep learning and pooled				
LSTM	0.481	0.811	0.401	0.737
Transformer	0.862	1.453	0.539	0.989
Pooled-LGBM	0.349	0.588	0.260	0.477
F. Forecast combinations				
Comb-Mean	0.403	0.678	0.347	0.637
Comb-InvMSE	0.331	0.557	0.272	0.500

Table 10. Volatility timing, one-day horizon. Average over the 25 instruments of the annualized Sharpe ratio, annualized return and daily turnover of the strategy (23) net of 5 basis points; average Sharpe ratio at 0, 10 and 20 basis points; and average annualized utility gain (24) relative to HAR-based timing for $\gamma = 5$, in basis points. The highest Sharpe ratio is in bold.

	Net of 5 bps			Sharpe at cost (bps)			
Model	Sharpe	Ann. ret. (%)	Turnover	0	10	20	Utility gain vs HAR (bps)
A. GARCH family							
GARCH	0.35	4.7	0.033	0.39	0.32	0.24	17
GJR	0.35	4.7	0.031	0.39	0.32	0.25	10
EGARCH	0.34	4.6	0.041	0.39	0.30	0.21	3
RealGARCH	0.33	4.5	0.062	0.40	0.26	0.13	-13
B. HAR family							
HAR	0.34	4.3	0.054	0.40	0.27	0.15	—
HAR-CJ	0.34	4.4	0.058	0.40	0.27	0.14	-1
SHAR	0.31	4.0	0.078	0.39	0.22	0.05	-42
HARQ	0.31	4.5	0.087	0.40	0.21	0.03	-79
LogHAR	0.32	4.3	0.073	0.40	0.24	0.07	-31
D. Machine learning — extended information set							
Ridge-X	0.30	4.1	0.085	0.40	0.21	0.01	-47
LASSO-X	0.31	4.2	0.082	0.40	0.21	0.03	-41
RF-X	0.31	4.3	0.078	0.39	0.22	0.05	-52
XGBoost-X	0.28	4.1	0.106	0.39	0.17	-0.06	-90
LightGBM-X	0.26	3.9	0.118	0.39	0.13	-0.12	-121
E. Deep learning and pooled							
LSTM	**0.38**	4.9	0.012	0.39	0.36	0.33	33
Transformer	0.35	4.7	0.004	0.35	0.34	0.33	-19
Pooled-LGBM	0.29	4.2	0.105	0.41	0.18	-0.05	-69
F. Forecast combinations							
Comb-Mean	0.34	4.5	0.052	0.40	0.28	0.16	10
Comb-InvMSE	0.33	4.3	0.056	0.39	0.26	0.13	-5

Table 11. One-day Value-at-Risk backtests. Mean violation rate across the 25 instruments and share of instruments on which the model passes the Kupiec unconditional-coverage and the Christoffersen conditional-coverage (CC) tests at the 5% level. Variance forecasts are rescaled to close-to-close return units with a trailing 250-day ratio.

Model	VaR 1%			VaR 5%		
	Viol. (%)	Kupiec pass (%)	CC pass (%)	Viol. (%)	Kupiec pass (%)	CC pass (%)
A. GARCH family						
GARCH	1.89	12	12	5.26	80	84
GJR	1.89	8	8	5.25	80	84
EGARCH	2.21	4	8	5.62	72	56
RealGARCH	1.79	24	24	5.18	92	88
B. HAR family						
HAR	1.73	12	20	5.04	96	84
HAR-CJ	1.71	20	28	5.08	96	80
SHAR	1.79	16	28	5.22	76	76
HARQ	1.94	8	12	5.33	76	84
LogHAR	1.79	20	20	5.16	96	92
D. Machine learning — extended information set						
Ridge-X	1.79	16	28	5.16	96	96
LASSO-X	1.77	16	32	5.14	100	96
RF-X	1.68	24	40	5.06	96	88
XGBoost-X	1.84	16	24	5.24	92	88
LightGBM-X	1.93	12	16	5.37	88	84
E. Deep learning and pooled						
LSTM	1.87	16	12	5.13	80	56
Transformer	1.97	16	12	5.14	88	44
Pooled-LGBM	1.90	4	20	5.38	84	84
F. Forecast combinations						
Comb-Mean	1.62	32	40	4.90	100	88
Comb-InvMSE	1.64	24	40	4.97	100	92

Table 12. Mean QLIKE, $h = 1$, by instrument and headline model. Common evaluation sample per instrument (December 2016 or later, to July 2026). Bold marks the lowest loss in the row.

Ticker	HAR	LogHAR	RealGARCH	LightGBM-X	Pooled-LGBM	LSTM	Comb-InvMSE
Equities/ETFs							
AAPL	0.182	0.164	0.165	0.216	0.175	0.309	**0.162**
GLD	0.208	0.203	0.204	0.251	0.204	0.435	**0.198**
IWM	0.439	0.149	0.148	0.190	0.156	0.244	**0.146**
JPM	0.162	0.139	0.140	0.163	0.144	0.207	**0.133**
MSFT	0.150	0.139	0.139	0.164	0.146	0.245	**0.136**
NVDA	0.155	0.152	0.153	0.175	0.157	0.254	**0.146**
QQQ	0.230	0.195	0.197	0.241	0.202	0.335	**0.185**
SPY	0.249	0.222	0.223	0.258	0.233	0.398	**0.210**
TLT	0.229	0.143	**0.142**	0.207	0.146	0.366	0.144
XOM	0.112	0.108	0.110	0.137	0.116	0.275	**0.104**
FX							
AUDUSD	0.143	0.134	0.136	0.145	0.130	0.300	**0.127**
EURUSD	0.128	0.121	0.123	0.137	0.121	0.295	**0.119**
GBPUSD	0.128	0.119	0.120	0.132	0.119	0.205	**0.114**
USDCAD	0.122	0.119	0.120	0.135	0.123	0.269	**0.114**
USDJPY	0.271	**0.229**	0.234	0.257	0.254	0.519	0.229
Crypto							
ADA	0.542	0.485	0.603	**0.430**	0.459	0.527	0.432
BTC	0.899	**0.272**	0.283	0.346	0.316	0.511	0.282
ETH	0.317	0.263	0.271	**0.259**	0.260	0.443	0.261
SOL	0.249	0.215	0.220	0.213	**0.183**	0.319	0.212
XRP	0.605	0.468	0.474	0.496	0.475	0.506	**0.447**
Futures							
CL	0.200	0.137	0.149	0.175	0.146	0.219	**0.132**
ES	0.219	0.197	0.200	0.220	0.191	0.328	**0.178**
GC	0.131	0.131	0.134	0.153	0.130	0.372	**0.127**
NQ	0.194	0.185	0.189	0.210	0.184	0.302	**0.166**
ZN	0.141	0.129	0.130	0.152	0.130	0.328	**0.126**